*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 6118–6126
Marseille, 11–16 May 2020

# RedDust: a Large Reusable Dataset of Reddit User Traits

**Anna Tigunova, Andrew Yates, Paramita Mirza, Gerhard Weikum**
Max Planck Institute for Informatics
Saarland Informatics Campus, Saarbrücken, Germany
{tigunova, ayates, paramita, weikum}@mpi-inf.mpg.de

### Abstract

Social media is a rich source of assertions about personal traits, such as *I am a doctor* or *my hobby is playing tennis.* Precisely identifying explicit assertions is difficult, though, because of the users' highly varied vocabulary and language expressions. Identifying personal traits from implicit assertions like *I've been at work treating patients all day* is even more challenging. This paper presents *RedDust*, a large-scale annotated resource for user profiling for over 300k Reddit users across five attributes: *profession*, *hobby*, *family status*, *age*, and *gender*. We construct RedDust using a diverse set of high-precision patterns and demonstrate its use as a resource for developing learning models to deal with implicit assertions. RedDust consists of users' personal traits, which are (attribute, value) pairs, along with users' post ids, which may be used to retrieve the posts from a publicly available crawl or from the Reddit API. We discuss the construction of the resource and show interesting statistics and insights into the data. We also compare different classifiers, which can be learned from RedDust. To the best of our knowledge, RedDust is the first annotated language resource about Reddit users at large scale. We envision further use cases of RedDust for providing background knowledge about user traits, to enhance personalized search and recommendation as well as conversational agents.

**Keywords:** personal knowledge, user profiling, conversational text, online forums

## 1 Introduction

Reddit is a popular social media platform for discussing a wide range of topics. It is an important source of information for data analysis on social media as it provides rich structure, abundance of data and covers a broad range of topics. Reddit is used by approximately 330 million users[1] with 2.8 million comments written each day[2]. Alexa.com ranks it as the 21st most popular website worldwide.

Despite its popular and rich data, few have considered Reddit as a source of data about users' personal traits like their professions and hobbies. Prior work has focused on Reddit as a source of demographic information, whereas we consider rich attributes like profession and hobbies in addition to demographic ones (age, gender, family status). Such data has many applications, including personalizing healthcare (Gyrard et al., 2018), recommendations, search, and conversational agents.

We address this gap by creating a labeled dataset of Reddit users (including their posts and comments) that covers five user attributes: *profession, hobby, family status, age,* and *gender*. We leveraged three high-precision approaches to identify predicates and their object values in users' posts: *(1)* natural language patterns matching assertions like *I am a flight attendant*, *(2)* bracket patterns matching structured assertions of users' ages and genders (*I [35m] just broke up with my girlfriend*), and *(3)* flair metadata specific to particular subfora. We used human judgments to validate the high-precision nature of these approaches before performing an analysis of the resulting dataset. To the best of our knowledge, RedDust is the first large scale semantic resource about user traits.

We illustrate the dataset's utility in two different use cases: building lexicons specific to attribute values and predicting users' attribute values expressed implicitly (e.g., *I've been fixing sinks all day*) after removing explicit assertions. This work makes the following contributions:

- We create a dataset of Reddit users traits, which are mined from users' personal assertions with several high-precision techniques. This resource is available at `https://pkb.mpi-inf.mpg.de/reddust`
- We perform a thorough analysis of the dataset, which sheds light on its structure and composition.
- We demonstrate two use cases for the dataset by building attribute-value-specific lexicons and performing classification of the labeled attributes with several state-of-the-art models.

## 2 Related work

**User Profiling in Online Communication**: The popularity of social media and online forums brings about massive amounts of user-generated content that is freely accessible. This has opened many research opportunities on text analysis, in particular on automatically identifying latent demographic features of online users for personalized downstream applications such as personalized search or recommendation. Such latent demographic attributes include *age* and *gender* (Basile et al., 2017; Bayot and Gonçalves, 2018; Burger et al., 2011; Fabian et al., 2015; Flekova et al., 2016a; Kim et al., 2017; Rao et al., 2010; Sap et al., 2014; Schwartz et al., 2013a; Vijayaraghavan et al., 2017), *personality* (Gjurković and Šnajder, 2018; Schwartz et al., 2013a), *regional origin* (Fabian et al., 2015; Rao et al., 2010), *political orientation* and *ethnicity* (Pennacchiotti and Popescu, 2011; Preoţiuc-Pietro et al., 2017; Preoţiuc-Pietro and Ungar, 2018; Rao et al., 2010; Vijayaraghavan et al., 2017), as well as *occupational class* mapped to income (Flekova et al., 2016b; Preoţiuc-Pietro et al., 2015).

[1] `https://redditblog.com/2018/11/13/holiday-on-reddit/`

[2] `https://www.digitaltrends.com/social-media/reddit-ads-promoted-posts/`

Most prior works on automatically identifying users' latent attributes from online communication rely on classification over hand-crafted features such as word/character n-grams (Basile et al., 2017; Burger et al., 2011; Rao et al., 2010), Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001) categories (Gjurković and Šnajder, 2018; Preoţiuc-Pietro et al., 2017; Preoţiuc-Pietro and Ungar, 2018), topic distributions (Flekova et al., 2016a; Pennacchiotti and Popescu, 2011; Preoţiuc-Pietro et al., 2015) and sentiment/emotion labels of words derived from existing emotion lexicon (Gjurković and Šnajder, 2018; Pennacchiotti and Popescu, 2011; Preoţiuc-Pietro et al., 2017; Preoţiuc-Pietro and Ungar, 2018). The recently prominent neural network approaches have also been adopted to solve the task (Bayot and Gonçalves, 2018; Kim et al., 2017; Tigunova et al., 2019; Vijayaraghavan et al., 2017). Among those prior works, Preoţiuc-Pietro et al. (2015), Basile et al. (2017), Bayot and Gonçalves (2018) and Tigunova et al. (2019) are the ones that infer users' latent attributes based solely on user-generated text, without relying on features specific to social media like hashtags or users' profile description.

**Dataset for User Profiling:** Automatic methods, particularly supervised learning approaches, for identifying users' personal attributes require a collection of user-generated content labelled with personal attributes of interest. Most of existing works mentioned above focus on user-generated content from Twitter, with a few exceptions that explore Facebook (Sap et al., 2014; Schwartz et al., 2013a) or Reddit (Fabian et al., 2015; Finlay, 2014; Gjurković and Šnajder, 2018) posts.

Data collection was mostly done via: manual annotation after a focused search with specific keywords or hashtags (Preoţiuc-Pietro et al., 2015; Rao et al., 2010), public profile linked to Twitter profile description (Burger et al., 2011; Flekova et al., 2016a), self-reports as part of an online survey (Finlay, 2014; Flekova et al., 2016a; Preoţiuc-Pietro et al., 2017; Preoţiuc-Pietro and Ungar, 2018; Sap et al., 2014; Schwartz et al., 2013b), or pattern-based extraction approach (e.g., `(I|i) (am|'m|was) born in + number (1920-2013)`) on user profile description or user posts (Fabian et al., 2015; Kim et al., 2017; Sloan et al., 2015; Tigunova et al., 2019). Several works (Basile et al., 2017; Bayot and Gonçalves, 2018) made use of labelled datasets published within the shared task on *author profiling* organized by the CLEF PAN lab (Francisco Manuel et al., 2017; Potthast et al., 2017).

There has been less effort on identifying demographic attributes of Reddit users compared with the body of work that exists for Twitter users, although Reddit posts have been exploited for other purposes such as determining *users' personality* (Gjurković and Šnajder, 2018), *mental health condition* (Cohan et al., 2018), *domestic abuse* (Schrading et al., 2015) and *irony detection* (Wallace et al., 2014), among others. Thelwall and Stuart (2019) investigate how the topic of subreddit influences the gender ratio within it. The study was performed on 100 subreddits grouped by interest, gender information about the users were collected by guessing it from their usernames, which is arguably a low-precision strategy. Smaller scale Reddit datasets exist for *gender*, *age* and *location* attributes (Fabian et al., 2015; Finlay, 2014), which are unfortunately not publicly available. As far as we know, we are the first to consider *hobby* as a personal attribute of interest to be identified from online communication.

## 3 Background

Reddit[3] is a social news website and forum where registered members can submit content including links, text posts, and images, which are then voted up or down by other members. Before elaborating on the creation of our dataset derived from Reddit posts, we describe several concepts on Reddit that are relevant for the data collection process.

**Posts and Comments:** Discussions on Reddit are organized in threads, which are initiated by an original *post* and may contain *comments* replying to the post and to other comment. This creates a hierarchical structure that resembles a conversation between users. Both posts and comments can be a textual content, a link with anchor text or images.

**Subreddits:** Reddit is organized into subreddits, which are fora that focus on specific topics. Those can be split by interest (sports, politics, etc), by country or community, type of content (text, gifs, videos), and so on. Subreddits have their own rules, but any registered user can create them. By convention, subreddits are prefixed with `/r`. For example, users discuss hockey in the `/r/hockey` subreddit.

**Flairs:** Flair is user or post metadata that is a unique feature of Reddit. Flair is generally a small image with a short text description that is attached to a post or a username. Flairs can be defined differently for specific purposes by each subreddit. For example, in `/r/travel` subreddit they may indicate the *country* of the user, *gender* in `/r/AskMen` and `/r/AskWomen` or users' *favorite teams* in `/r/hockey`. Flairs for posts can be useful to filter and search for a particular content.

## 4 Dataset Creation

*RedDust* is a dataset containing a collection of Reddit users, or *redditors*. Each redditor in the dataset is associated with posts and comments they produce, and personal attributes resulting from pattern-based mining approach as well as flairs. In this work, we consider five personal attributes including *gender*, *age*, *family status*, *profession* and *hobby*. The dataset is created from the openly published Reddit dump[4], which spans between 2006 and 2018.

There are several criteria on which users and posts/comments that are included in RedDust, i.e., users who posted between 10 and 100 posts/comments, and posts/comments containing between 20 and 100 terms after filtering. We filtered out hyperlinks and user mentions (i.e., `@nickname`) from the original posts/comments.

Some subreddits are likely to contain many false positives, such as those concerned with video games or role playing. This leads to personal assertions talking about the users' projected persona in a particular context (e.g., *I am a priest*

[3] https://www.reddit.com/

[4] https://files.pushshift.io/reddit/

*looking for a guild*). To mitigate this source of false positives, we blacklisted subreddits about gaming, fantasy, and virtual reality from the top 500 subreddits sorted by number of unique users. Posts made to blacklisted subreddits were discarded. Similarly, we do discard posts that contain quotations in order to reduce the possibility of the user referring to a third person (*... and he shouted "Hands on the counter, I am a cop!"*).

For attributes that usually have a unique value (i.e., *gender*, *age* and *family status*) we also exclude users who state multiple different values to avoid introducing false positives. Meanwhile, we allow each user to have multiple attribute values for *profession* and *hobby*. The age of a given user is calculated relative to his or her age when writing the most recent comment.

In the following subsections we discuss particular techniques used to extract values or labels for each personal attribute.

## 4.1 Gender

Gender has been the most popular user attribute to predict in existing user profiling work, particularly on Reddit (Fabian et al., 2015; Thelwall and Stuart, 2019; Vasilev, 2018). In RedDust, we consider gender as a binary predicate (*female* or *male*) as has been done in prior work.

Instead of weak supervision like considering usernames as a means for gender classification, as was done by Thelwall and Stuart (2019), we look for self-reported gender assertions, which provide labels of higher precision. Specifically, we identified users' gender using the following methods:

- **Natural language patterns**. Following Fabian et al. (2015), we manually created a set of patterns that indicate a specific gender. They have the general form of `(I am|I'm) a? <gender indicator>`, meaning that matches should contain *I am* or *I'm*, optionally followed by an article *a*, then a word that indicates gender like *man* or *mother*. A comprehensive list of patterns we used is given in Table 1, and the indicative gender words are shown in Table 2. Although the gender of a given user can be expressed in a longer snippet like *I am a great mother*, we do not allow extra words like *great* to appear before gender-indicating words. This reduces false positives from statements like *I'm a far cry from a mother*.

  Still, there are a variety of false positive patterns, which are tricky to avoid. Those could be imaginary situations (*I dreamed I am a mother*), reported speech (*she said "I am a mother"*–we don't consider the sentences with quotes for this purpose) and some others (*I am a boy scout*). Those are hard to avoid automatically, and thus, manual examination over the extracted posts is necessary.

- **Bracket patterns**. In certain situations, users often volunteer to indicate their demographic information in order to give their posts more context (*I [30f] was dating this guy [35m]...*). This is common in relationship-related subreddits where the users' age and gender are often relevant to the discussions. These cues are generally written in round or square brackets. To reduce false positives, we do not consider such patterns when they appear without brackets. To capture gender and age expressed in this way, we look for patterns of the form `(I|I'm|me) [<number>(m|f)]`.

- **Flairs**. Like Vasilev (2018), we also consider gender-indicating flairs attached to users. This logic is subreddit-specific, so we restrict ourselves to common subreddits. For example, in subreddits `/r/AskWomen` and `/r/AskMen` the flair is one of *male, female, trans*, and so on, whereas in `/r/tall` and `/r/short` the flair is either *pink, blue*, or *other*.

## 4.2 Age

We label users' posts with age predicate using similar techniques as for gender:

- **Natural language patterns**. To infer users' age, we utilized five patterns listed in Table 1, with pattern (v) specifically designed to avoid false positives as in *I am 6 feet tall*. We then calculated the exact age for patterns (i)-(iii) by subtracting the birth year from the publishing year of the post containing such patterns.

- **Bracket patterns**. Numbers indicating age were jointly collected along with gender, as described in the above-mentioned bracket patterns for gender.

Finally, we made sure that the obtained ages for users in RedDust are within the range of 10-100 years old, since users under 13 are not allowed to register and there are unlikely to be many users above 100 years old. This is helpful for reducing false positives, such as those in conditional sentences (*as if I am 5 years old*).

## 4.3 Family status

We consider family status as a binary predicate indicating whether a person is *single* or has a *partner*. As for gender, we relied on *natural language patterns* containing *indicative words*, which are detailed in Table 1 and 2, respectively. We distinguished two cases of indicative words: (i) `self-status indicator`, used when the speaker refers to her own status (*I am 'divorced'*); and (ii) `partner indicator`, when the speaker refers to the existence of a partner (*My 'boyfriend'*).

We additionally collected matches of negated patterns of both (i) and (ii), i.e., `I am not <self-status indicator>` and `I don't have a <partner indicator>`, in order to expand the labelled data. Furthermore, given that the indicator word *single* often used in a more general context (e.g., *single player*, *single bed*), we restricted the patterns containing this particular word that it should be immediately followed by punctuation, conjunctions or few allowed words like *father*.

## 4.4 Profession

To obtain profession labels we consulted a list of occupation names from Wikipedia[5] and recursively added all titles

[5] en.wikipedia.org/wiki/Category:Lists_of_occupations

| **attribute** | **pattern(s)** |
|---|---|
| gender | `(I am\|I'm) a? <gender indicator>` (e.g., *man*, *mother*) |
| age | (i) `I (was\|am) born in <four digit year>`<br>(ii) `I (was\|am) born in <two digit year>`<br>(iii) `I was born on <day, month, year>`<br>(iv) `I am <number> years old`<br>(v) `I am <number>` immediately followed by punctuation or conjunctions (*and*, *but*, etc.) |
| family status | (i) `I am <self-status indicator>` (e.g., *divorced*, *single*)<br>(ii) `(my\|I have a) <partner indicator>` (e.g., *wife*, *boyfriend*) |
| profession | `(I am\|I'm) a <profession name>` |
| hobby | `<phrase indicator>` (e.g., *I enjoy*, *I like*) `<hobby name>` |

Table 1: Patterns for labeling Reddit users with personal attributes.

| **attribute** | **value** | `word/phrase indicators` |
|---|---|---|
| gender | female<br>male | *woman*, *female*, *girl*, *lady*, *wife*, *mother*, *sister*<br>*man*, *male*, *boy*, *husband*, *father*, *brother* |
| family status | single<br>partner | `self-status`: *single*, *divorced*, *widow*, *spouseless*, *celibate*, *unmarried*, *unwed*, *fancy-free*<br>`self-status`: *married*, *engaged*, *dating*<br>`partner`: *boyfriend*, *spouse*, *girlfriend*, *fiancee*, *lover*, *partner*, *wife*, *husband* |
| hobby | - | *my hobby is*, *I am/I'm fond of*, *I am/I'm keen on*, *I like*, *I enjoy*, *I go in for*, *I take joy in*, *I adore, I love*, *I play, I fancy*, *I am/I'm a fan of, I am/I'm fascinated by*, *I am/I'm interested in*, *I appreciate, I practise*, *I am/I'm mad about* |

Table 2: Words and phrases considered as indicators used in patterns for labeling personal attributes.

under subcategories. The resulting list consists of about 1K professions and contains a lot of fine grained occupations, some of which are redundant or ambiguous. Our strategy is to capture as many profession assertions as possible, giving users of RedDust the opportunity to filter and group the professions depending on their specific use cases.

Each profession in the list was considered as `profession name` in the pattern `(I am|I'm) a <profession name>` that we used to label Reddit users with the *profession* attribute. After performing pattern matching against the whole Reddit dataset, we were left with 832 unique profession names in RedDust.

### 4.5 Hobby

Similar to the profession attribute, we obtained a list of hobbies from Wikipedia[6] and utilized them as `hobby name` in our natural language patterns for the *hobby* attribute. We used a diverse set of patterns of the form `<phrase indicator> <hobby name>`, where `phrase indicator` is a phrase like *my hobby is* or *I enjoy*, as listed in Table 2. Using the pattern matching approach, users in RedDust were labeled with 336 unique hobby names in total.

### 4.6 Labeling evaluation

To validate the high-precision nature of our labeling approach, we asked three human annotators to verify the correctness of labels for each predicate. We randomly sampled 50 labeled posts for each attribute and asked annotators to indicate whether the given label matched the user's actual assertion. The decision to accept or reject the label was based on a majority vote from the annotators.

| **attribute** | **precision** | **#false positives** | **#disagreements** |
|---|---|---|---|
| gender | 0.96 | 2 | 2 |
| age | 1.0 | 0 | 2 |
| family status | 0.86 | 7 | 8 |
| profession | 0.96 | 2 | 2 |
| hobby | 0.94 | 3 | 9 |
| avg/total | 0.94 | 14 | 23 |

Table 3: Number of false positives and inter-rater agreement for each attribute.

The results of this human evaluation are shown in Table 3. In total there were 23 instances without perfect annotator agreement (out of 250 total instances for five attributes), which indicated 14 false positives after taking a majority vote. Half of these false positives came from the family status attribute, due to ambiguous usage of words like *partner* in statements like *I have a partner in crime for that*. Despite such false positives, the average labeling precision for all personal attributes in RedDust is 94%. Furthermore, we also measured annotator agreement with Fleiss' Kappa as 0.67 in average for all attributes, which is a substantial agreement; with the worst agreement (0.59) reached for the family status attribute.

[6] en.wikipedia.org/wiki/List_of_hobbies

| attribute | #users | #posts | #subreddits |
|---|---|---|---|
| gender | 54.88K | 2.49M | 28.25K |
| age | 122.20K | 5.80M | 44.07K |
| family status | 11.77K | 0.56M | 14.76K |
| profession | 74.86K | 3.63M | 37.49K |
| hobby | 89.07K | 4.42M | 41.31K |
| total | 352.78K | 16.9M | 165.88K |

Table 4: Overall RedDust statistics.

### 4.7 Privacy and licensing

We note that our dataset consists of attribute-value pairs for real subjects (Reddit users) who may desire to edit or delete their posts. We took several steps in order to protect users' privacy and to preserve their ability to remove their posts. First, we do not disclose any usernames or user IDs to represent each labeled user in RedDust, so that users retain the ability to sever links to their posts (e.g., by *removing author information* of a certain post without deleting the post itself). Second, we represent the posts or comments belonging to each labeled user with the corresponding Reddit post/comment IDs, which are revealed in the Reddit URLs. Hence, users are still able to opt to exclude certain posts/comments from our dataset.

Regarding licensing and usage, the homepage of our project page[7] where we published the dataset provides data usage information and states that the resource is available under a *Creative Commons* license (CC BY 4.0).

## 5 Data statistics and analysis

In this section we present the quantitative and qualitative analysis of the RedDust resource. We present in Table 4 the overall statistics of the dataset. Further statistics on specific attribute values can be found in the RedDust distribution.

In Figure 1, we plot the chart of the user count per each post count within the admissible 20-100. From this plot we conclude that the users in our dataset tend to have a small number of posts.

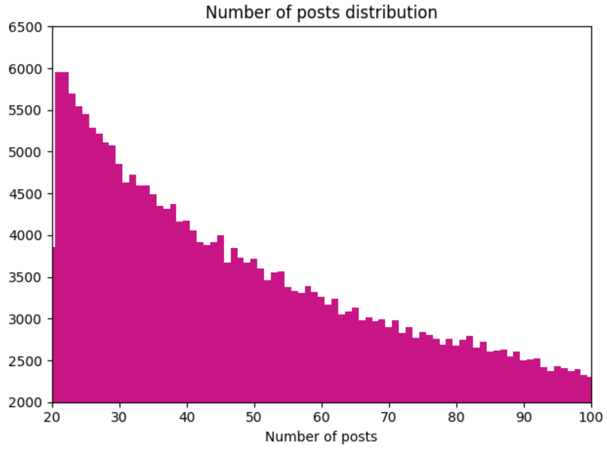

Figure 1: Counts of users having $x$ number of posts.

As previously mentioned, the profession attribute has 832 possible values in total. However, just about 5% of professions have over 2000 users, which we consider to be a sufficient number to train predictive models for automatically labeling users with personal attributes (see Section 6).

### 5.1 Analysis of multiple personal attributes

Almost 19K users in RedDust have two personal attributes known, three for 980 and four for 28, which amount to 6% of the users having multiple personal attributes in total. For such users it is interesting to look at interplay between different personal traits, for instance, the correlation between users' occupations and general interests. We plot a heat map in Figure 2 which represents the co-occurring values for these two predicates. For this experiment as well as the subsequent ones, we limit the number of professions and hobbies to the top $k$ ones ($k = 20$ and $k = 30$ for profession and hobby, respectively), sorted by the number of labeled users per value.

We observed intuitive correlations such as: *musicians* often play *guitar*; *runners* have *running* as the main interest; *college students* like to *read* but are also interested in *video games* five times as much as any other professions; and curiously, *shooting* is popular among *photographers*, most probably because of *shooting* being an ambiguous term.

We also considered other pairs of attributes, namely profession and gender, for which we show the gender distribution of each profession in Figure 3. The analysis revealed common prejudices like *female nannies* or *male programmers*, as well as several surprising insights (prevalence of *female runners* and *bartenders*) possibly specific to Reddit communities.

### 5.2 Subreddit analysis

For each labeled user in our dataset, we retrieved posts authored by the user along with the subreddits these posts were written in, which allows us to investigate the dependencies between personal attributes and subreddits. For instance, Figure 4 shows the distribution of genders in RedDust as well as the distribution of most popular subreddits for each gender.

The chart exposes *females* as the dominant users in our dataset, which does not conform to statistics from other sources.[8] However, due to peculiarities of our labeling patterns, note that RedDust does not give the true outlook of real life population as well as Reddit demography. We also observed that the ranked lists of most popular subreddits for either gender are highly similar, with the exception of `/r/AskWomen` and `/r/TwoXChromosomes` subreddits, which specifically target a certain gender. However, subreddits like `/r/AskReddit` and `/r/relationships` are generally popular across various demographics. To further investigate the correlations between professions and subreddits, we decided to exclude such *popular subreddits* from our analysis, which we define as subreddits in which at least 15 users have posted their comments.

For each profession $P$, we calculated the score for each subreddit $S$ as the probability of users labeled with $P$ will

[7] `https://pkb.mpi-inf.mpg.de/reddust`

[8] For example, `techjunkie.com/demographics-reddit/`.

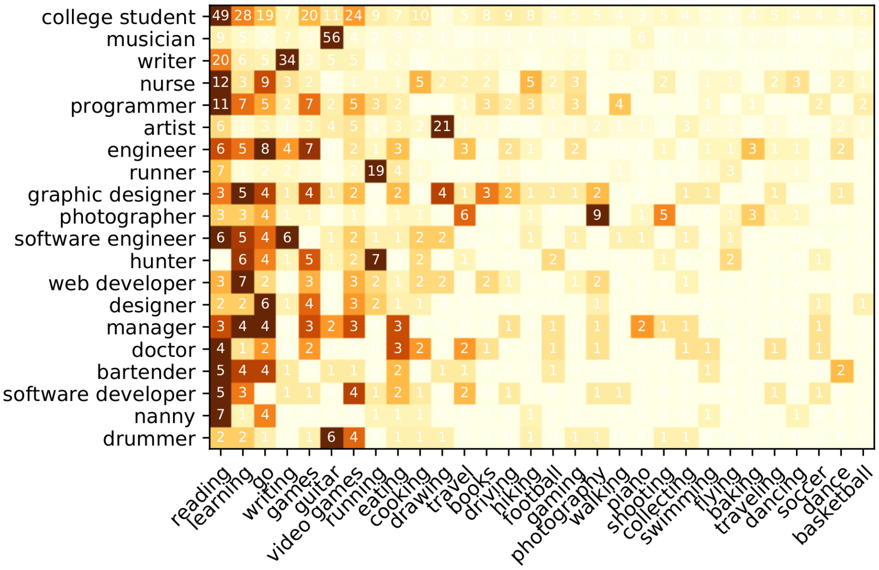

Figure 2: Co-occurrence of the most common professions and hobbies.

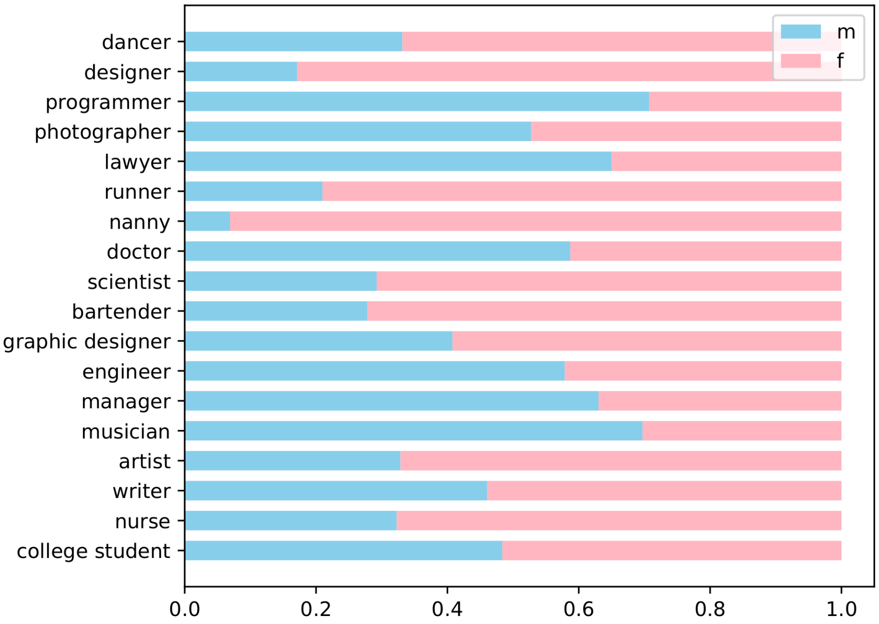

Figure 3: Gender distribution among professions.

| nurse | photographer | web developer |
|---|---|---|
| `/r/nursing` | `/r/photography` | `/r/webdev` |
| `/r/MakeupAddiction` | `/r/photocritique` | `/r/web_design` |
| `/r/loseit` | `/r/itookapicture` | `/r/forhire` |
| `/r/AskWomen` | `/r/Instagram` | `/r/Entrepreneur` |
| `/r/StudentNurse` | `/r/analog` | `/r/programming` |

Table 5: Typical subreddits for *nurse*, *photographer* and *web developer*.

post/comment in $S$, i.e., $Pr(S|prof = P)$. Table 5 showcases the top scoring subreddits for selected professions, which are remarkably relevant to a given profession and are also not likely found for other professions.

| | **gender** | | **family status** | |
|---|---|---|---|---|
| | **AUROC** | **acc** | **AUROC** | **acc** |
| LogReg | 0.49 | 0.57 | 0.50 | 0.50 |
| MLP | 0.49 | 0.57 | 0.63 | 0.63 |
| CNN | 0.88 | 0.80 | **0.90** | 0.81 |
| HAM$_{\text{CNN-attn}}$ | **0.91** | **0.86** | 0.89 | **0.82** |

Table 6: Evaluation results for binary attributes.

## 6 Use Cases

In this section we discuss potential applications of RedDust. For example, RedDust may be used both as a tool for the analysis and creation of trait-specific lexicons as well as a labeled dataset for training predictive models, among other use cases.

|  | age |  |  | profession |  |  | hobby |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  | **MFR** | **AUROC** | **nDCG** | **MFR** | **AUROC** | **nDCG** | **MFR** | **AUROC** | **nDCG** |
| LogReg | 1.0 | 0.78 | 0.90 | **14.6** | 0.74 | 0.34 | **19.1** | 0.79 | 0.44 |
| MLP | 1.0 | 0.82 | **0.96** | 12.9 | 0.78 | 0.35 | 17.3 | **0.80** | **0.46** |
| CNN | **2.8** | 0.85 | 0.95 | 12.7 | 0.83 | 0.45 | 18.6 | 0.77 | 0.30 |
| HAM$_{\text{CNN-attn}}$ | 2.4 | **0.88** | **0.96** | 14.5 | **0.85** | **0.47** | 18.3 | **0.80** | 0.30 |

Table 7: Evaluation results for multi-valued attributes.

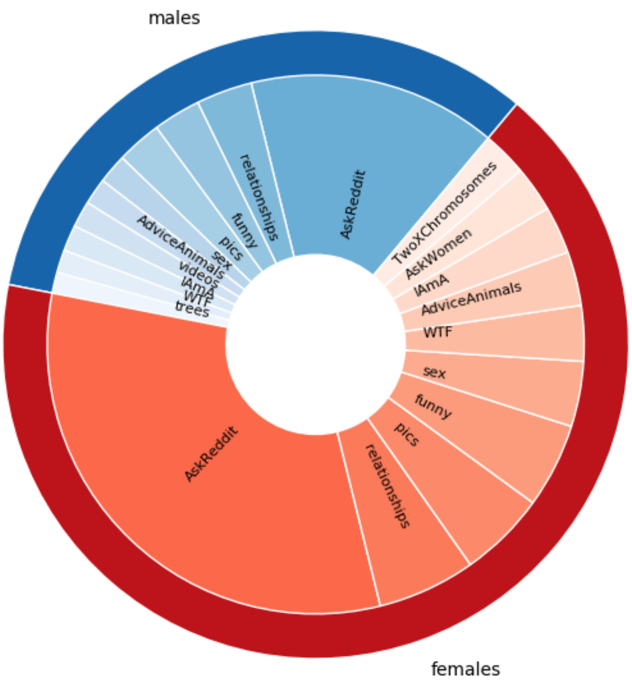

Figure 4: Most popular subreddits for genders

## 6.1 Trait-specific characteristic terms

RedDust may be used to build *lexicons* specific to different personal *user traits*. Here we use “traits” to refer to attribute-value pairs, such as “profession-doctor” or “hobby-swimming”. Such lexicon can later be used for personalization in many downstream applications (e.g., targeted advertisements based on the trait-specific search terms). For that purpose, we counted the number of word occurrences for each attribute value, excluding *common* and *rare* words (occurring more than 15 times and less than 4 times, respectively). We also removed proper names, mentions of trait attribute (”profession”) and values (”writer”) and phrases.

In Table 8, we list the resulting typical words for several professions found by RedDust (ordered by frequency). We can observe that the RedDust lexicons reflect the terminologies uttered by people having specific professions. For example, 33% of people who ever wrote any of the words from [*bartend, limes, bitters, tippers, dreamcast, cognac*] on Reddit were indeed bartenders, and similarly, 20% for musicians.

## 6.2 Learning predictive models

For the second use case of RedDust, we utilized the labeled data to train predictive models for identifying users’ personal attributes given their posts/utterances, which is useful for (i) expanding labeled personal traits of Reddit users for a more comprehensive analysis of Reddit community and (ii) downstream applications like personalized recommendation or conversational agents. However, we assume that in natural conversations or discussions, people rarely reveal their personal attributes via explicit assertions like *I am a man*, and hence, require the models to learn to pick up implicit cues. To better simulate a prediction task in such a situation, we exclude all posts that explicitly mention attribute values, i.e., the ones used to label users in RedDust. We then build binary classifiers for *gender* and *family status* and multi-class ones for the other attributes to predict attribute values given users’ posts as input.

### 6.21. Experimental setup

We conducted the experiments by splitting the dataset into a 9:1 proportion for training and test, respectively. Each instance for the classification models is a user represented by all the user’s posts, each of which is a bag of words. Following prior work (Tigunova et al., 2019), we remove punctuation, stopwords, usernames, hash tags and hyperlinks from the posts. Input words are represented using *word2vec embeddings* (Mikolov et al., 2013) trained on GoogleNews. For each personal attribute considered, we built and evaluated a classification model using several methods:

**Logistic regression** (LogReg). We built a multinomial logistic regression classifier to get the probability for each attribute value. The input to the classifier is the average embeddings of all words in all posts of a given user.

**Multilayer Perceptron** (MLP). We applied a shallow MLP classifier with one hidden layer of size 100 and ReLU activation, which was trained for 200 epochs. The input to the MLP is the same as for logistic regression.

**Convolutional Neural Network** (CNN). We leveraged the CNN architecture proposed by Bayot and Gonçalves (2018) for predicting the *age* and *gender* of Twitter users. The model takes as input the concatenation of all words from a given user. We used filters of sizes 2 and 3 to produce feature maps, which are then concatenated and classified with a fully-connected layer. We applied 64 convolutional filters and trained the model for 100 epochs.

**Hidden Attribute Models** (HAM). We used a publicly available implementation[9] of a strong personal attribute classification method from Tigunova et al. (2019). HAM$_{\text{CNN-attn}}$ is a neural model that utilizes a CNN with attention mechanism. This model considers users’ posts as input in a hierarchical way: first, the representations of

[9] https://github.com/Anna146/HiddenAttributeModels/

|  | **writer** | **bartender** | **musician** | **web developer** | **lawyer** |
|---|---|---|---|---|---|
| RedDust lexicon | wordpress | bartend | percussion | fonts | statue |
|  | spacebar | limes | composing | stackoverflow | litigation |
|  | erotica | bitters | triangles | serif | counsel |
|  | screenwriter | tippers | concerto | sharply | prosecutor |
|  | goblins | dreamcast | lydian | freelancing | plea |
|  | scandalous | cognac | amaj | joomla | defendant |

Table 8: Typical words for *writer*, *bartender*, *musician*, *web developer* and *lawyer*.

words are put through a CNN for each post separately to create the post's latent representation. Then, the model applies attention mechanism on post representations, to find the importance weights of each post. The weighted average of all posts is then passed on to a fully-connected layer to yield the probability of each attribute value. We used the following hyperparameters: 128 filters of size 2 and attention layer of size 150; the model was trained for 70 epoch.

For MLP, CNN and HAM, we chose hyperparameters including number of epochs by performing grid search using cross-validation on the training set.

### 6.22. Evaluation metrics

**Binary attributes**. We computed *accuracy* and *area under the curve (AUROC)* metrics to evaluate the classifiers for *gender* and *family* status attributes.

**Multi-valued attributes**. For the *age* attribute, we put age numbers into buckets corresponding to different age categories to be predicted, namely: (a) 13-23: *teenager*, (b) 24-45: *adult*, (c) 46-65: *middle-aged* and (d) 66-100: *senior*; assuming that children under 13 are not allowed to use Reddit. For *profession* and *hobby* attributes, we considered only a subset of attribute values for which we have sufficient number of users, i.e., 250. We further refined the list of values manually by merging similar ones, like *police officer* and *cop*, which left us with 69 unique professions and 89 distinct hobbies.

For all multi-valued attributes, we reported model performance in terms of *AUROC* and two ranking metrics: *Mean First Relevant (MFR)* and *normalized Discounted Cumulative Gain (nDCG)*. To compute AUROC in the multi-class setting, we binarized the labels and computed one-vs-all scores. Ranked list of attribute values were obtained by ranking the predicted probabilities given by the model. Given the ranked list of attribute values, we computed MFR (Mean First Rank) as the average ranking of the first relevant label for all instances in the test set (Fuhr, 2018). For nDCG, we used binary labels (correct or incorrect value).

### 6.23. Experiment Results

The evaluation results for binary and multi-valued attributes are presented in Table 6 and Table 7, respectively. For *gender* and *family status*, the best performing method HAM$_{\text{CNN-attn}}$ yields high accuracy of 0.86 and 0.82, respectively, showing the promising utility of RedDust for building predictive models for binary attributes.

For multi-valued attributes (*age*, *profession* and *hobby*), we observed that the more attribute values the models have to predict, the more degraded the models' performance were. However, considering the vast amount of profession and hobby values for multi-class classification, the best performing methods CNN and HAM$_{\text{CNN-attn}}$ are still able to rank the first correct label within the top 20% of values on average. We hope that our dataset will facilitate research on such *extreme classification* (Varma, 2018) task, as well as on identifying implicit personal traits.

## 7 Conclusion

In this work we described RedDust, a large semantic resource about Reddit users' personal traits. Over 350,000 users were labeled with values for five personal attributes (*gender*, *age*, *family status*, *profession* and *hobby*) by using a combination of high-precision patterns to capture personal assertions and Reddit metadata. RedDust's utility was illustrated in two use cases leveraging the dataset to build trait-specific lexicons and to predict users' traits based on only implicit assertions.

## 8 Bibliographical References

Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., and Nissim, M. (2017). N-GrAM: New Groningen Author-profiling Model—Notebook for PAN at CLEF 2017. In *Working Notes Papers of the CLEF 2017 Evaluation Labs*.

Bayot, R. K. and Gonçalves, T. (2018). Age and gender classification of tweets using convolutional neural networks. In *Machine Learning, Optimization, and Big Data*, Cham. Springer International Publishing.

Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating gender on twitter. In *Proceedings of EMNLP'11*, July.

Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., and Goharian, N. (2018). "SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions". In *Proceedings of the 27th International Conference on Computational Linguistics*.

Fabian, B., Baumann, A., and Keil, M. (2015). Privacy on reddit? towards large-scale user classification. In *Proceedings of ECIS'15*.

Finlay, S. C. (2014). Age and gender in reddit commenting and success.

Flekova, L., Carpenter, J., Giorgi, S., Ungar, L., and Preoţiuc-Pietro, D. (2016a). Analyzing biases in human perception of user age and gender from text. In *Proceedings of ACL'16 (Volume 1: Long Papers)*.

Flekova, L., Preoţiuc-Pietro, D., and Ungar, L. (2016b). Exploring stylistic variation with age and income on twitter. In *Proceedings of ACL'16 (Volume 2: Short Papers)*.

Francisco Manuel, Rangel Pardo, Rosso, P., Potthast, M., and Stein, B. (2017). Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In *Working Notes Papers of the CLEF 2017 Evaluation Labs*.

Fuhr, N. (2018). Some common mistakes in ir evaluation, and how they can be avoided. In *ACM SIGIR Forum*, volume 51, pages 32–41. ACM.

Gjurković, M. and Šnajder, J. (2018). Reddit: A gold mine for personality prediction. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, NAACL-HLT'18*.

Gyrard, A., Gaur, M., Shekarpour, S., Thirunarayan, K., and Sheth, A. (2018). Personalized health knowledge graph. In *First International Workshop on Contextualized Knowledge Graphs*.

Kim, S. M., Xu, Q., Qu, L., Wan, S., and Paris, C. (2017). Demographic inference on twitter using recursive neural networks. In *Proceedings of ACL'17 (Volume 2: Short Papers)*, July.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Pennacchiotti, M. and Popescu, A.-M. (2011). A machine learning approach to twitter user classification. In *Proceedings of ICWSM'11*.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71.

Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., and Stein, B. (2017). Overview of PAN'17: Author Identification, Author Profiling, and Author Obfuscation. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 7th International Conference of the CLEF Initiative (CLEF 17)*.

Preoţiuc-Pietro, D., Lampos, V., and Aletras, N. (2015). An analysis of the user occupational class through twitter content. In *Proceedings of ACL/IJCNLP'15 (Volume 1: Long Papers)*, July.

Preoţiuc-Pietro, D. and Ungar, L. (2018). User-level race and ethnicity predictors from twitter text. In *Proceedings of COLING'18*.

Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., and Ungar, L. (2017). Beyond binary labels: Political ideology prediction of twitter users. In *Proceedings of ACL'17 (Volume 1: Long Papers)*.

Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. (2010). Classifying latent user attributes in twitter. In *Proceedings of SMUC'10*.

Sap, M., Park, G., Eichstaedt, J., Kern, M., Stillwell, D., Kosinski, M., Ungar, L., and Schwartz, H. A. (2014). Developing age and gender predictive lexica over social media. In *Proceedings EMNLP'14*, October.

Schrading, N., Alm, C. O., Ptucha, R., and Homan, C. (2015). An analysis of domestic abuse discourse on reddit. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2577–2583.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., and Ungar, L. H. (2013a). Personality, gender, and age in the language of social media: The open-vocabulary approach. In *PloS one*.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., et al. (2013b). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Sloan, L., Morgan, J., Burnap, P., and Williams, M. (2015). Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *PloS one*, 10(3):e0115545.

Thelwall, M. and Stuart, E. (2019). She's reddit: A source of statistically significant gendered interest information? *Information Processing Management*, 56(4):1543 – 1558.

Tigunova, A., Yates, A., Mirza, P., and Weikum, G. (2019). Listening between the lines: Learning personal attributes from conversations. In *The Web Conference*. ACM.

Varma, M. (2018). Extreme Classification: Tagging on Wikipedia, Recommendation on Amazon & Advertising on Bing. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 1897–1897, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Vasilev, E. (2018). Inferring gender of reddit users. Master's thesis, Universität Koblenz-Landau, Universitätsbibliothek.

Vijayaraghavan, P., Vosoughi, S., and Roy, D. (2017). Twitter demographic classification using deep multimodal multi-task learning. In *Proceedings of ACL'17 (Volume 2: Short Papers)*, July.

Wallace, B. C., Kertz, L., Charniak, E., et al. (2014). Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 512–516.