Tell Me Something I Don’t Know: Analyzing OkCupid Profiles

Juan Shishido[‡*], Jaya Narasimhan[¶†], Matar Haller[§†]

https://youtu.be/dtgmMj8W298

Abstract—In this paper, we present an analysis of 59,000 OkCupid user profiles that examines online self-presentation by combining natural language processing (NLP) with machine learning. We analyze word usage patterns by self-reported sex and drug usage status. In doing so, we review standard NLP techniques, cover several ways to represent text data, and explain topic modeling. We find that individuals in particular demographic groups self-present in consistent ways. Our results also suggest that users may unintentionally reveal demographic attributes in their online profiles.

Index Terms—natural language processing, machine learning, supervised learning, unsupervised learning, topic modeling, okcupid, online dating

Introduction

Online dating has become a common and acceptable way of finding mates. In the United States, 41 percent of adults know someone who uses online dating, 29 percent know someone who has met a partner this way, and 59 percent believe online dating is a good way to meet people [Pew16]. In 2015, online dating sites or mobile dating apps were used by 27 percent of 18-24 year olds, 22 percent of 25-34 year olds, 21 percent of 35-44 year olds, 13 percent of 45-54 year olds, and 12 percent of 55-64 year olds [Pew16]. Relative to 2013, usage across every age group, except 25-34 year olds, increased. Given the popularity of online dating, the way that people self-present online has broad implications for the relationships they pursue.

Previous studies suggest that the free-text portion of online dating profiles is an important factor (after photographs) for assessing attractiveness [Fio08]. The principle of homophily posits that people tend to associate and bond with individuals who are similar to themselves and that this strongly structures social networks, most prominently by race and ethnicity [McP01]. Perhaps not surprisingly, research suggests that homophily extends to online dating, with people seeking mates similar to themselves [Fio05]. However, it remains unclear whether people within particular demographic groups, such as sex or ethnicity, self-present in similar ways when searching for a mate online.

In this paper, we analyze demographic trends in online self-presentation. Specifically, we focus on whether people signal demographic characteristics through the way they present themselves online. We extend previous natural language processing analyses of online dating [Nag09] by using a much larger sample[1] and by combining NLP with supervised and unsupervised machine learning. We leverage multiple approaches including clustering and topic modeling as well as feature selection and modeling strategies. By exploring the relationships between free-text self-descriptions and demographics, we discover that we can predict users’ demographic makeup and also find some unexpected insights into unintentional signaling of demographic characteristics.

Code and data for this work are available in our `okcupid` GitHub repository[2]. A Jupyter notebook with the analysis results is also available[3].

Data

Description

Profile information[4] was available for 59,946 OkCupid users that were members as of 06/26/2012, lived within 25 miles of San Francisco, had been active in the previous year, and had at least one photo in their profile [Wet15]. The data set contained free-text responses to 10 essay prompts as well as the following user characteristics: age, body type, diet, drinking status, drug usage status, education level, ethnicity, height, income, job type, location, number of children, sexual orientation, attitude toward pets, religion, sex, astrological sign, smoking status, number of language spoken, and relationship status.

This public[5] data set was selected for its diverse set of essay prompts and availability of detailed user characteristics, which enabled us to examine the connection between online self-presentation and demographics. This data set has previously been used to demonstrate the basics of text analysis as well as how to fit a simple logistic regression model to predict sex using only height

* *Corresponding author: juanshishido@berkeley.edu*
‡ *School of Information, University of California, Berkeley*
† *These authors contributed equally.*
¶ *Department of Electrical Engineering and Computer Science, University of California, Berkeley*
§ *Helen Wills Neuroscience Institute, University of California, Berkeley*

[1] [Nag09]’s uses a sample of 1,000 individuals.

[2] https://github.com/juanshishido/okcupid.

[3] https://github.com/juanshishido/okcupid/blob/master/OkNLP-paper.ipynb

[4] https://github.com/rudeboybert/JSE_OkCupid. Our original data source was Everett Wetchler’s `okcupid` repository (https://github.com/everett-wetchler/okcupid). However, after commit `0d62e62`, in which the data was "fully anonimized" to exclude essays, we switched to Kim’s repository. Kim uses the original Wetchler data.

[Kim15]. The present study extends previous work by analyzing additional features and by introducing novel analyses.

Preprocessing

Line break characters, URLs, and HTML tags were removed from the essay text. Multiple periods, dashes, and white spaces were replaced by single instances, and all text was converted to lowercase. Essays were segmented, first into sentences and then into individual terms, using spaCy's [Hon16][6] default tokenizer, which is well suited for online communication as it maintains emoticons as discrete tokens. This allowed us to differentiate between the syntactic way that special characters are traditionally used and the meaning that's conveyed when they are used in particular combinations. Punctuation was removed *after* the text was tokenized[7]. Finally, users who wrote less than five words for a given essay were removed from the analysis.

In order to reduce the number of categories, we combined drug usage status levels. Specifically, users who responded "sometimes" or "often" were grouped into a "yes" category. Individuals who answered "never" were assigned to the "no" group and we created an "unknown" category for users who did not answer.

Methods

Term Frequency-Inverse Document Frequency

Machine learning tasks require numerical inputs. There are several ways to represent text as numerical feature vectors. Features typically correspond to distinct tokens or to sequences of adjacent tokens. A token is a series of characters, such as a word, that is treated as a distinct unit [Bir10].

One way to represent a corpus, or collection of text documents, is as a matrix of token counts. This weights terms by their absolute frequencies. Often, highly-weighted terms, such as "a" or "the," are not informative, so token counts are weighted using term frequency-inverse document frequency (tf-idf).

Tf-idf is the product of the term frequency and the inverse document frequency. The term frequency refers to the *relative* frequency of term t in document d. The inverse document frequency is the log of the total number of documents N to the number of documents that contain term t.

Log-Odds-Ratio

One metric for comparing word usage across groups is to calculate the log-odds-ratio. The odds for word w in the usage of group g are defined as $O_{iw} = \frac{f_{iw}}{(1-f_{iw})}$ where f_{iw} is the frequency count of word w normalized by total count of words used by group i. If a word is used only by one group, its log-odds-ratio is infinite. Therefore, a constant is added to each frequency when calculating the odds. The log of the ratio of the adjusted odds between groups can then be used to compare word usage across groups.

Non-negative Matrix Factorization

For document clustering, the document corpus is projected onto a k-dimensional semantic space, with each axis corresponding to a particular topic and each document being represented as a linear combination of those topics [Xu_03]. Methods such as latent semantic indexing require the derived latent semantic space to be orthogonal, so this class of methods does not work well when corpus topics overlap, as is often the case. Conversely, non-negative matrix factorization (NMF) does not require the latent semantic space to be orthogonal, and therefore is able to find directions for related or overlapping topics.

NMF was applied to each essay of interest using scikit-learn [Ped11][8], which uses the coordinate descent solver. NMF utilizes document frequency counts, so the tf-idf matrix for unigrams, bigrams, and trigrams was calculated, while limiting tokens to those appearing in at least 0.5 percent of the documents. NMF was calculated with k dimensions, which factorized the tf-idf matrix into two matrices, W and H. The dimensions were `n_samples x k` and `k x n_features` for W and H, respectively. Group descriptions were given by top-ranked terms in the columns of H. Document membership weights were given by the rows of W. The maximum value in each row of W determined essay group membership.

Permutation Testing

Permutation tests provide an exact sampling distribution of a test statistic under the null hypothesis [Ger12] by computing the test statistic for every manner by which labels can be associated with the observed data. In practice, permutations are rarely ever completely enumerated. Instead, the sampling distribution is approximated by randomly shuffling the labels P times.

The likelihood of the observed test statistic is determined as the proportion of times that the absolute value of the permuted test statistics are greater than or equal to the absolute value of the observed test statistic. This is the p-value for a two-tailed hypothesis. Permutation-based methods can be used to compare two samples or to assess the performance of classifiers [Oja10].

There are several advantages to using randomization to make inferences as opposed to parametric methods. Permutation tests do not assume normality, do not require large samples, and "can be applied to all sorts of outcomes, including counts, durations, or ranks" [Ger12].

Approach

Our analyses focused on two demographic dimensions — sex and drug usage — and on two essays — "My self summary" and "Favorite books, movies, shows, music, food." These essays were selected because they were answered by most users. "The most private thing I am willing to admit" prompt, for example, was ignored by 32 percent of users.

We began by exploring the lexical features of the text as a way to determine whether there were differences in writing styles by demographic group. We considered essay length, the use of profanity and slang terms, and part-of-speech usage.

Essay length was determined based on the tokenized essays. A list of profane words was obtained from the "Comprehensive Perl Archive Network" website. Slang terms include words such as "dough," which refers to money, and acronyms like "LOL." These terms come from the Wiktionary Category:Slang page[9]. Note that there is overlap between the profane and slang lists.

[5] As authorized by OkCupid president and co-founder Christian Rudder [Kim15].

[6] We used version 0.101.0. GitHub, 10 May 2016. https://github.com/spacy-io/spaCy/releases/tag/0.101.0.

[7] Punctuation is needed for the sentence tokenizer and sentences are important for the part-of-speech tagging.

[8] We used version 0.17.1. GitHub, 18 Feb 2016. https://github.com/scikit-learn/scikit-learn/releases/tag/0.17.1-1. This is particularly important for NMF as the coordinate descent solver is the default as of 0.17.0. Using the deprecated projected gradient solver will lead to different results.

[9] https://simple.wiktionary.org/wiki/Category:Slang.

Each token in the corpus was associated with a lexical category using spaCy's part-of-speech tagger. spaCy supports 19 coarse-grained tags[10] that expand upon Petrov, Das, and McDonald's universal part-of-speech tagset [Pet11].

Differences in lexical features by demographic were analyzed using permutation testing. We first compared average essay length by sex. Next, we examined whether the proportion of females using profanity was different than the proportion of males using such terms. The same was done for slang words. Finally, we compared the average proportion of adjectives, nouns, and verbs and identified the most distinctive terms in each lexical category by sex using the smoothed log-odds-ratio, which accounts for variance.

We also analyzed text semantics by transforming the corpus into a tf-idf matrix using spaCy's default tokenizer. We chose to include unigrams, bigrams, and trigrams[11]. Stop words[12] and terms that appeared in less than 0.5 percent of documents were removed. Stemming, the process of removing word affixes, was not performed. This resulted in a vocabulary size of 2,058 for the self-summaries essay and 2,898 for the favorites essay.

Non-negative matrix factorization was used to identify latent structure in the text. This structure represented "topics" or "clusters" which were described by particular tokens. In order to determine whether particular demographic groups were more likely to write about certain topics, the relative distribution of users over topics was plotted. In cases where we were able to create superordinate groupings from NMF topics — for example, by combining semantically similar clusters — we used the log-odds-ratio to find their distinctive tokens.

Based on our findings, we decided to fit a logistic regression model to predict drug usage status.

Results

In this section, we describe our lexical- and semantic-based findings.

We first compared lexical-based characteristics on the self-summary text by sex. Our sample included 21,321 females and 31,637 males[13]. On average, females wrote significantly longer essays than males (150 terms compared to 139, $p < 0.001$).

Next, we compared the proportion of users who utilized profanity and slang. Profanity was rarely used in the self-summary essay. Overall, only 6 percent of users included profane terms in their self-descriptions. The difference by sex was not statistically significant (5.8% of females versus 6.1% of males, $p = 0.14$).

Not surprisingly, slang was much more prevalent than profanity. 56 percent of users used some form of slang in their self-summary essays and females used slang at a significantly lower rate than males (54% versus 57%, $p < 0.001$).

To compare part-of-speech usage, we first associated part-of-speech tags with every token in the self-summary corpus. This resulted in counts by user and part-of-speech. Each user's counts were then normalized by the user's essay length to account for essay length differences between users. Of the 19 possible part-of-speech tags, we focused on adjectives, nouns, and verbs. The proportions of part-of-speech terms used is shown in Table 1.

Part-of-Speech	Female	Male
Adjectives ⋆⋆	10.61%	10.16%
Nouns ⋆⋆	18.65%	18.86%
Verbs	18.28%	18.27%

***TABLE 1:** Proportion of part-of-speech terms used, by sex. Asterisks (⋆⋆) denote statistically significant differences at the 0.001 level.*

Part-of-Speech	Female	Male
Adjectives	independent sweet my sassy silly happy warm favorite girly fabulous	nice cool its that few interesting martial most masculine more
Nouns	girl family who yoga men gal heels love dancing friends	guy computer engineer guitar sports software women video technology geek
Verbs	love am laugh laughing dancing adore loving dance appreciate being	m was play playing laid 'll working hit moved been

***TABLE 2:** The 10 most-distinctive adjective, noun, and verb tokens , by sex.*

Females used significantly more adjectives than males, while males used significantly more nouns than females ($p < 0.001$ for both). There was no difference in verb usage between the sexes ($p = 0.91$).

In addition to part-of-speech usage, we explored specific terms associated with parts-of-speech that were distinctive to a particular sex. We did this using the log-odds-ratio. Table 2 summarizes this, below.

Distinctly-female adjectives are mostly descriptive. Males, on the other hand, use more quantity-based and demonstrative adjectives. For nouns, females focus on relationship- and experience-based terms while males write about work, sports, and technology. (Note that `m` corresponds to the contracted form of "am" when "Im" (no apostrophe) is tokenized and that `'ll` is the contracted form of "will" in terms such as "I'll.")

NMF was then used to provide insight into the underlying topics that users chose to use to describe themselves. Selecting the number of NMF components (topics to which users are clustered) is an arbitrary and iterative process. For the self-summary essay, we chose 25 components, which resulted in a diverse, but manageable, set of topics.

Several expected themes emerged. Many users chose to highlight personality traits, for example "humor" or "easy-going," while others focused on describing the types of activities they enjoyed. Hiking, traveling, and cooking were popular choices. Others chose to mention what kind of interaction they were seeking, whether that was a long-term relationship, a friendship, or sex. Topics and the highest weighted tokens for each are summarized in Table 3. Note that topic names were hand-labeled.

In order to determine whether there were differences in the topics that OkCupid users chose to write about in their self-summaries, we plotted the distribution over topics by demographic split. This allowed us to identify if specific topics were distinct to

[10] https://spacy.io/docs#token-postags.

[11] Unigrams are single tokens. Bigrams refer to two adjacent and trigrams to three adjacent tokens.

[12] Stop words are words that appear with very high frequency, such as "the" or "to."

[13] The difference between the number of users in the data set and the number of users in the analysis is due to the fact that we drop users that write less than five tokens for a particular essay.

Topic	Tokens
meet & greet	meet new people, looking meet new, love meeting new, new friends, enjoy meeting, interesting people, want meet, ’m new, people love, experiences
the city	san francisco, moved san francisco, city, living san francisco, just moved san, native, san diego, grew, originally, recently
enthusiastic	love travel, love laugh, love outdoors, love love, laugh, dance, love cook, especially, life love, love life
straight talk	know, just, want, ask, message, just ask, really, talk, write, questions
about me	’m pretty, ’m really, ’m looking, ’m just, say ’m, think ’m, ’m good, ’m trying, nerd, ’m working
novelty	new things, trying new, trying new things, new places, learning new things, exploring, restaurants, things love, love trying, different
seeking	’m looking, guy, relationship, looking meet, share, woman, nice, just looking, man, partner
carefree	easy going, ’m easy going, easy going guy, pretty easy going, laid, love going, enjoy going, simple, friendly, likes
casual	guy, lol, chill, nice, old, pretty, alot, laid, kinda, wanna
enjoy	like, ’d like, things like, really like, n’t like, feel like, stuff, like people, like going, watch
transplant	moved, sf, years ago, school, east coast, city, just moved, college, went, california
nots	n’t, ca n’t, does n’t, really, wo n’t, n’t like, n’t know, n’t really, did n’t, probably
moments	spend time, good time, lot, free time, spending time, lot time, spend lot, time friends, time ’m, working
personality	humor, good sense humor, good time, good conversation, sarcastic, love good, dry, good company, appreciate, listener
amusing	fun loving, ’m fun, having fun, outgoing, guy, girl, adventurous, like fun, looking fun, spontaneous
review	let ’s, think, way, self, right, thing, say, little, profile, summary
region	bay area, moved bay area, bay area native, grew, living, ’m bay area, east bay, raised bay area, east, originally
career-focused	work hard, play hard, hard working, progress, harder, job, try, love work, company, busy
locals	born, raised, born raised, california, raised bay area, college, school, sf, berkeley, oakland
unconstrained	open minded, creative, honest, relationship, adventurous, curious, passionate, intelligent, heart, independent
active	enjoy, friends, family, hiking, watching, outdoors, traveling, hanging, cooking, sports
creative	music, art, live, movies, live music, play, food, games, dancing, books
carpe diem	live, world, fullest, enjoy life, experiences, passionate, love life, moment, living life, life short
cheerful	person, people, make, laugh, think, funny, kind, happy, honest, smile
jet setter	’ve, lived, years, world, traveled, year, spent, countries, different, europe

TABLE 3: *Self-summary topics and associated terms.*

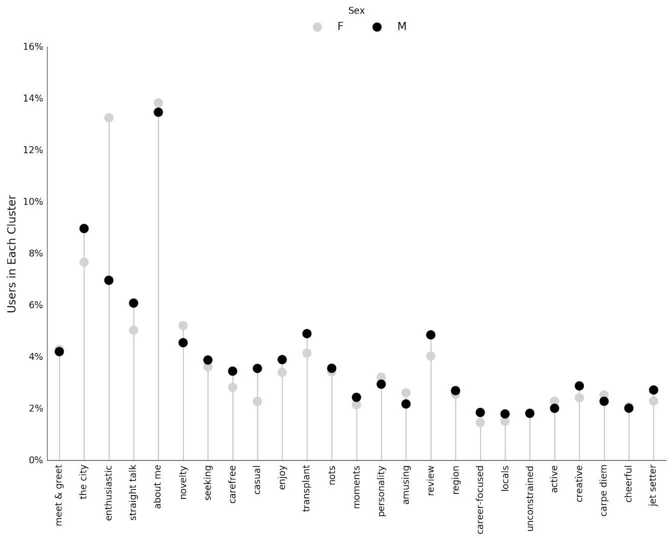

Fig. 1: *Self-summary distribution over topics*

particular demographic groups.

Figure 1 shows the distribution over topics by sex for the self-summary essay. The highest proportion of users, of either sex, were in the "about me" topic. This is not surprising given the essay prompt. For most topics, females and males were mostly evenly distributed. For example, the proportion of females who emphasized their careers or travel or other topics was similar to the proportion of males who did the same. One exception was with the "enthusiastic" topic, to which females belonged at almost twice the rate of males. Users in this group used modifiers such as, "love," "really," and "absolutely" regardless of the activities they were describing.

We further examined online self-presentation by considering the other available essays in the OkCupid data set. Previous psychology research suggests that a person’s preferred music styles are tied to their personalities [Col15], and it is possible that this extends to other media, such as books or movies. We next analyzed the "Favorite books, movies, shows, music, food" essay.

As with the self-summaries, we removed users who wrote less than 5 tokens for this essay (11,836 such cases). Note that because the favorites text is less expository and more list-like, we did not perform a lexical-based analysis. Instead, we used NMF to identify topics (or genres). Like with the self-summaries, we chose 25 topics. Table 4 lists the topics and a selection of their highest weighted tokens.

The topics for this essay were less distinctive than the topics for the self-summaries. In some cases, genres (or media) overlapped. For example, the "TV-comedies-0" group included "The Walking Dead," which is a drama. There was also overlap between groups. Still, we decided to keep 25 components. The granularity these topics provided was used for further analyses. We created superordinate groupings from the topics from which we extracted distinctive tokens for particular demographic groups, showing the approach’s flexibility. Figure 2 shows the distribution over topics, by sex.

The most popular topics, for both females and males, were "TV-hits" and "music-rock," with about 16 percent of each sex writing about shows or artists in those groups. We found more separation between the sexes in the favorites essay than we did with the self-summaries. As with the self-summary essay,

Topic	Tokens
like	like, music like, movies like, really like, stuff, food like, things, like music, books like, like movies
TV-hits	mad men, arrested development, breaking bad, 30 rock, tv, parks, sunny, wire, dexter, office
enthusiastic	love food, love music, love movies, love love, cook, love good, eat, food, love read, books love
favorite-0	favorite, favorite food, favorite movies, favorite books, favorite music, favorite movie, favorite book, favorite shows, favorite tv, time favorite
genres-movies	sci fi, action, comedy, horror, fantasy, movies, drama, romantic, classic, adventure
genres-music	hip hop, rock, r&b, jazz, reggae, rap, pop, country, classic, old
misc-0	fan, reading, food 'm, right, 'm big, really, currently, music 'm, just, open
TV-comedies-0	big bang theory, met mother, big lebowski, friends, house, office, community, walking dead, new girl, bones
genres-food	italian, thai, mexican, food, indian, chinese, japanese, sushi, french, vietnamese
nots	ca n't, watch, n't really, does, n't like, does n't, think, eat, n't watch tv, n't read
teen	harry potter, hunger games, twilight, dragon tattoo, pride prejudice, harry met sally, disney, vampire, trilogy, lady gaga
everything	books, movies, food, music, shows, country, dance, action, lots, horror
movies-drama-0	eternal sunshine, spotless mind, litte miss sunshine, amelie, garden state, lost, life, beautiful, lost translation, beauty
time periods	80, let, good, 90, life, just, 70, world, time, man
avid	read lot, time, watch, listen, recently, lately, love read, watch lot, favorites, just read
misc-1	list, just, long, ask, way, goes, things, try, favorites, far
music-rock	david, black, john, tom, radiohead, bob, brothers, beatles, black keys, bowie
movies-sci-fi	star, lord, wars, rings, star trek, trilogy, series, matrix, princess, bride
TV-comedies-1	modern family, family guy, office, south park, met mother, glee, simpsons, american dad, 30 rock, colbert
movies-drama-1	fight club, shawshank redemption, pulp fiction, fear loathing, peppers, red hot, vegas, american, catcher rye, big lebowski
kinds	kinds music, love kinds, kinds food, kinds movies, listen, different, country, foods, comedy, action
favorite-1	favorite book, favorite movie, food, music, good, fav, book read, reading, great, best
novelty	enjoy, new, types, trying, reading, things, foods, types music, films, different
TV-drama	game thrones, ender 's game, walking dead, true blood, series, currently, hunger games, dexter, song ice, boardwalk empire
genres-books	fiction, non fiction, science fiction, fiction books, read non fiction, historical fiction, films, books, documentaries, biographies

TABLE 4: *Favorites topics and associated terms.*

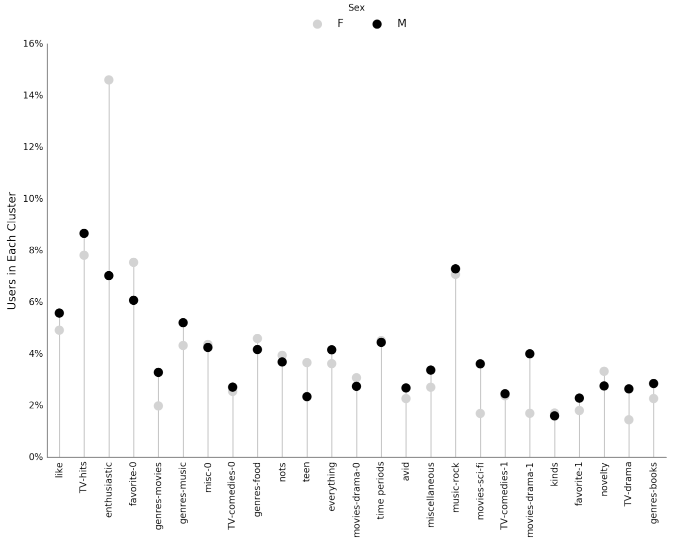

Fig. 2: *Favorites distribution over topics, by sex*

the enthusiastic group was distinctly female. A distinctly male category included films such as "Fight Club" and "The Shawshank Redemption" and musicians such as the Red Hot Chili Peppers.

We created superordinate groupings by combining clusters. There were four groups related to movies. In order to extract demographic-distinctive tokens, we used the smoothed log-odds-ratio which accounts for variance as described by Monroe, Colaresi, and Quinn [Mon09]. The top movies for females were Harry Potter, Pride & Prejudice, and Hunger Games while males favored Star Wars, The Matrix, and Fight Club. The "movies-sci-fi" and "movies-drama-1" groups, whose highest weighted tokens referred to the male-favored movies, had a higher proportion of males than females. Similarly, the "teen" group, which which corresponded to female-favored movies, had a higher proportion of females. This reflects the terms found by the log-odds-ratio.

Figure 3 shows the distribution over topics by drug usage. In this demographic category, users self-identified as drug users or non-drug users. To this, we added a third level for users who declined the state their drug usage status. There were 6,859 drug users, 29,402 non-drug users, and 11,849 users who did not state their drug usage status ("unknown").

There was more intra-cluster variation in the distribution of users across topics than for the demographic split by sex. Interestingly, the distribution across topics of users for whom we had no drug usage information — those in the "unknown" category — tended to track the distribution of self-identified drug users. In other words, the proportion of drugs users and unknown users in most topics was similar. This was especially true in cases where difference in proportions of drug users and non-drug users was large. This unexpected finding may suggest that individuals who did not respond to the drug usage question abstained in order to avoid admitting they did use drugs.

Although we were unable to test this hypothesis directly due to lack of the true drug-usage status for these users, the manner by which free-text writing styles may unintentionally disclose demographic attributes is an intriguing avenue for research. We used a predictive modeling approach to attempt to gain insights into this question. Specifically, we trained a logistic regression model on a binary outcome, using only drug users and non-

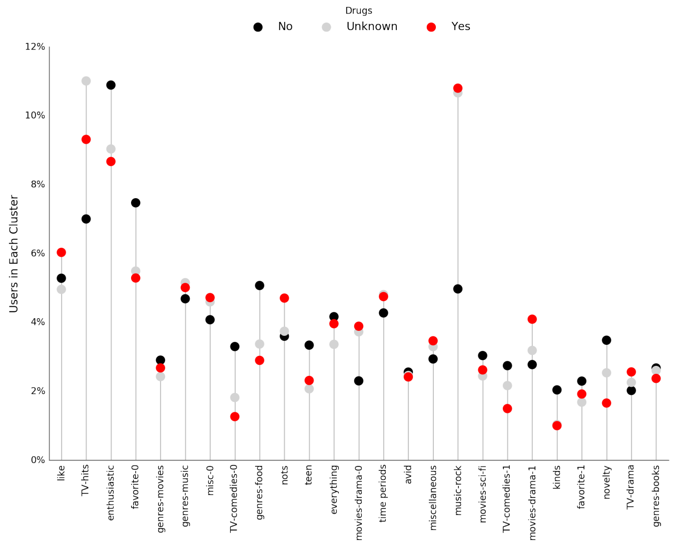

***Fig. 3:** Favorites distribution over topics, by drug usage status*

drug users. We used tf-idf weights on unigrams, bigrams, and trigrams as in the previous analyses. We also balanced the classes by randomly sampling 6,859 accounts from the non-drug user population. The data was split into training (80%) and test (20%) sets in order to assess model accuracy. We then predicted class labels on the group of unknown drug usage status.

Our initial model, which used only the "Favorites" essay text, accurately predicted 68.0 percent of drug users. When applied to the unknown users upon which the model was not trained, the model predicted that 55 percent of the unknown users were drug users and that 45 percent were not. When we examined the proportion of predicted user by NMF cluster, however, we found intriguing patterns. In the "music-rock" group — the group with the largest disparity between users and non-users — 84 percent of unknowns were classified as drug users. In contrast, only 25 percent of the unknowns in the "TV-comedies-0" group were classified as such. While this cluster included "The Big Lebowski," which is identified as a "stoner film" [She13], it also features "The Big Bang Theory," "How I Met Your Mother," "NCIS," "New Girl," and "Seinfeld," which we would argue are decidedly not drug-related.

These results prompted us examine if we could predict drug usage status based on text alone. For this, we combined the text of all 10 essays and dropped the 2,496 users who used less than five tokens in the full-text. As before, we randomly sampled from the non-users in order to balance the classes and split the data into training and test sets.

The full-text model accuracy increased to 72.7 percent. We used the feature weights to find the 25 most-predictive drug-usage terms. These are listed below, with the odds ratio[14] shown in parentheses.

```
sex (68.96), shit (45.51), music (20.95),
weed (18.46), party (15.54), beer (14.18),
dubstep (13.86), fuck (12.28), drinking (11.48),
smoking (11.39), partying (10.59), chill (9.45),
hair (8.84), park (8.09), fucking (7.93), dj (7.9),
burning (7.78), electronic (7.05), drunk (6.67),
ass (6.36), reggae (6.18), robbins (5.81),
dude (5.74), smoke (5.68), cat (5.5)
```

Drug users in this data set reference drinking, smoking, partying, and music more than non-users and also use particular profane terms.

Conclusion and Future Work

The current study extended previous NLP analyses of online dating profiles. The scope of this work was larger than previous studies, both because of the size of the data set and because of the novel combination of NLP with both supervised and unsupervised machine learning techniques, such as logistic regression and NMF. To our knowledge, there is currently no study that combines these techniques to identify unintentional cues in online self-presentation or uses them to predict demographics from free-text self descriptions. The idea that people may unintentionally be providing information about themselves in the way that they answer questions online is an intriguing avenue for future research and can also be extended to deception online.

This work serves as an initial exploration for analyzing self-presentation in the context of online dating. Given the availability of other demographic characteristics, such as ethnicity and education level, future work will focus on describing the ways in which other demographic groups tend to describe themselves. We would also like to explore recent advancements in language modeling techniques, such as word embeddings. Most importantly, future work will involve exploring methods to help us better identify deception. If the data ever becomes available, we would like to explore how the *way* that people choose to self-present affects the interactions they have.

Acknowledgements

This work began as a final project for the Applied Natural Language Processing course at the School of Information at the University of California, Berkeley. We would like to thank Marti Hearst for her guidance in the "right" way to do NLP and in pushing us to explore new and exciting data sets. We would also like to thank David Bamman for fruitful discussions on NLP and ideas for permutation testing. We would especially like to thank our reviewers, in particular David Lippa. His comments were invaluable for helping us organize our thoughts and analyses.

References

[Bir10] Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python. "O'Reilly Media, Inc.".

[Col15] Collingwood, J. (2015). Preferred Music Style Is Tied to Personality. Psych Central. Retrieved on June 22, 2016, from http://psychcentral.com/lib/preferred-music-style-is-tied-to-personality/

[Fio05] Fiore, A. T., & Donath, J. S. (2005, April). Homophily in online dating: when do you like someone like yourself?. In CHI'05 Extended Abstracts on Human Factors in Computing Systems (pp. 1371-1374). ACM.

[Fio08] Fiore, A. T., Taylor, L. S., Mendelsohn, G. A., & Hearst, M. (2008, April). Assessing attractiveness in online dating profiles. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 797-806). ACM.

[Ger12] Gerber, A. S., & Green, D. P. (2012). Field experiments: Design, analysis, and interpretation. WW Norton.

[Hon16] Honnibal, M (2016). spaCy. [Computer software]. https://spacy.io/.

[Kim15] Kim, A. Y., & Escobedo-Land, A. (2015). OkCupid Data for Introductory Statistics and Data Science Courses. Journal of Statistics Education, 23(2), n2.

[14] Logistic regression coefficient estimates are given as log-odds-ratios. The odds-ratios, which say how much a one unit increase affects the odds of being a drug user, are calculated by exponentiating.

[McP01] McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. Annual review of sociology, 415-444.

[Mon09] Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. Political Analysis, 16(4), 372-403.

[Nag09] Nagarajan, M., & Hearst, M. A. (2009, March). An Examination of Language Use in Online Dating Profiles. In ICWSM.

[Oja10] Ojala, M., & Garriga, G. C. (2010). Permutation tests for studying classifier performance. Journal of Machine Learning Research, 11(Jun), 1833-1863.

[Ped11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.

[Pet11] Petrov, S., Das, D., & McDonald, R. (2011). A universal part-of-speech tagset. arXiv preprint arXiv:1104.2086.

[Pew16] Smith, Aaron, & Anderson, Monica (2016). 5 Facts About Online Dating. Retrieved from http://www.pewresearch.org/fact-tank/2016/02/29/5-facts-about-online-dating/.

[She13] Sheffield, Rob (2013). 10 Best Stoner Movies of All Time. Rolling Stones. Retrieved on June 23, 2016, from http://www.rollingstone.com/movies/lists/the-greatest-stoner-movies-of-all-time-20130606

[Wet15] Everett Wetchler, okcupid, (2015), GitHub repository, https://github.com/everett-wetchler/okcupid.git

[Xu_03] Xu, W., Liu, X., & Gong, Y. (2003, July). Document clustering based on non-negative matrix factorization. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (pp. 267-273). ACM.