I'm a Belieber: Social Roles via Self-identification and Conceptual Attributes

Charley Beller, Rebecca Knowles, Craig Harman
Shane Bergsma[†], Margaret Mitchell[‡], Benjamin Van Durme
Human Language Technology Center of Excellence
Johns Hopkins University, Baltimore, MD USA
[†]University of Saskatchewan, Saskatoon, Saskatchewan Canada
[‡]Microsoft Research, Redmond, Washington USA
charleybeller@jhu.edu, rknowles@jhu.edu, craig@craigharman.net,
shane.a.bergsma@gmail.com, memitc@microsoft.com, vandurme@cs.jhu.edu

Abstract

Motivated by work predicting coarse-grained author categories in social media, such as gender or political preference, we explore whether Twitter contains information to support the prediction of *fine-grained* categories, or *social roles*. We find that the simple self-identification pattern "*I am a* __" supports significantly richer classification than previously explored, successfully retrieving a variety of fine-grained roles. For a given role (e.g., **writer**), we can further identify characteristic *attributes* using a simple possessive construction (e.g., *writer's* __). Tweets that incorporate the attribute terms in first person possessives (*my* __) are confirmed to be an indicator that the author holds the associated social role.

1 Introduction

With the rise of social media, researchers have sought to induce models for predicting latent *author attributes* such as gender, age, and political preferences (Garera and Yarowsky, 2009; Rao et al., 2010; Burger et al., 2011; Van Durme, 2012b; Zamal et al., 2012). Such models are clearly in line with the goals of both computational advertising (Wortman, 2008) and the growing area of computational social science (Conover et al., 2011; Nguyen et al., 2011; Paul and Dredze, 2011; Pennacchiotti and Popescu, 2011; Mohammad et al., 2013) where big data and computation supplement methods based on, e.g., direct human surveys. For example, Eisenstein et al. (2010) demonstrated a model that predicted where an author was located in order to analyze regional distinctions in communication. While some users explicitly share their GPS coordinates through their Twitter clients, having a larger collection of automatically identified users within a region was preferable even though the predictions for any given user were uncertain.

We show that media such as Twitter can support classification that is more fine-grained than gender or general location. Predicting *social roles* such as **doctor**, **teacher**, **vegetarian**, **christian**, may open the door to large-scale passive surveys of public discourse that dwarf what has been previously available to social scientists. For example, work on tracking the spread of flu infections across Twitter (Lamb et al., 2013) might be enhanced with a factor based on aggregate predictions of author occupation.

We present two studies showing that first-person social content (tweets) contains intuitive signals for such fine-grained roles. We argue that non-trivial classifiers may be constructed based purely on leveraging simple linguistic patterns. These baselines suggest a wide range of author categories to be explored further in future work.

Study 1 In the first study, we seek to determine whether such a signal exists in *self-identification*: we rely on variants of a single pattern, "*I am a* __", to bootstrap data for training balanced-class binary classifiers using unigrams observed in tweet content. As compared to prior research that required actively polling users for ground truth in order to construct predictive models for demographic information (Kosinski et al., 2013), we demonstrate that some users specify such properties publicly through direct natural language.

Many of the resultant models show intuitive strongly-weighted features, such as a **writer** being likely to tweet about a *story*, or an **athlete** discussing a *game*. This demonstrates self-identification as a viable signal in building predictive models of social roles.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 181–186,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

Role	Tweet
artist	I'm an Artist..... the last of a dying breed
belieber	@justinbieber I will support you in everything you do because I am a belieber please follow me I love you 30
vegetarian	So glad I'm a vegetarian.

Table 1: Examples of self-identifying tweets.

#	Role	#	Role	#	Role
29,924	little	5,694	man	564	champion
21,822	big	...	...	559	teacher
18,957	good	4,007	belieber	556	writer
13,069	huge	3,997	celebrity	556	awful
13,020	bit	3,737	virgin	...	...
12,816	fan	3,682	pretty	100	cashier
10,832	bad	...	...	100	bro
10,604	girl	2,915	woman	...	...
9,981	very	2,851	beast	10	linguist
...	...	...	...	...	...

Table 2: Number of self-identifying users per "role". While rich in interesting labels, cases such as *very* highlight the purposeful simplicity of the current approach.

Study 2 In the second study we exploit a complementary signal based on characteristic *conceptual attributes* of a social role, or concept class (Schubert, 2002; Almuhareb and Poesio, 2004; Paşca and Van Durme, 2008). We identify typical attributes of a given social role by collecting terms in the Google n-gram corpus that occur frequently in a possessive construction with that role. For example, with the role **doctor** we extract terms matching the simple pattern "*doctor's* __".

2 Self-identification

All role-representative users were drawn from the free public 1% sample of the Twitter Firehose, over the period 2011-2013, from the subset that selected English as their native language (85,387,204 unique users). To identify users of a particular role, we performed a case-agnostic search of variants of a single pattern: *I am a(n)* __, and *I'm a(n)* __, where all single tokens filling the slot were taken as evidence of the author self-reporting for the given "role". Example tweets can be seen in Table 1, examples of frequency per role in Table 2. This resulted in 63,858 unique roles identified, of which 44,260 appeared only once.[1]

We manually selected a set of roles for further exploration, aiming for a diverse sample across: occupation (e.g., **doctor**, **teacher**), family (**mother**), disposition (**pessimist**), religion (**christian**), and "followers" (**belieber**, **directioner**).[2] We filtered users via language ID (Bergsma et al., 2012) to better ensure English content.[3]

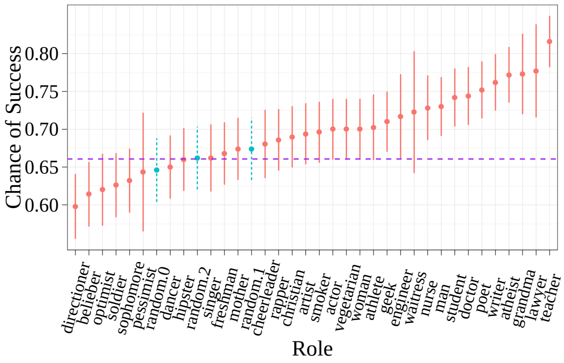

Figure 1: Success rate for querying a user. Random.0,1,2 are background draws from the population, with the mean of those three samples drawn horizontally. Tails capture 95% confidence intervals.

For each selected role, we randomly sampled up to 500 unique self-reporting users and then queried Twitter for up to 200 of their recent publicly posted tweets.[4] These tweets served as representative content for that role, with any tweet matching the self-reporting patterns filtered. Three sets of background populations were extracted based on randomly sampling users that self-reported English (post-filtered via LID).

Twitter users are empowered to at any time delete, rename or make private their accounts. Any given user taken to be representative based on a previously posted tweet may no longer be available to query on. As a hint of the sort of user studies one might explore given access to social role prediction, we see in Figure 1 a correlation between self-reported role and the chance of an account still being publicly visible, with roles such as **belieber** and **directioner** on the one hand, and **doctor** and **teacher** on the other.

The authors examined the self-identifying tweet of 20 random users per role. The accuracy of the self-identification pattern varied across roles and is attributable to various factors including quotes, e.g. *@StarTrek Jim, I'm a DOCTOR not a download!*. While these samples are small (and thus estimates of quality come with wide variance), it

[1] Future work should consider identifying multi-word role labels (e.g., *Doctor Who fan*, or *dog walker*).

[2] Those that follow the music/life of the singer Justin Bieber and the band One Direction, respectively.

[3] This removes users that selected English as their primary language, used a self-identification phrase, e.g. *I am a belieber*, but otherwise tended to communicate in non-English.

[4] Roughly half of the classes had less than 500 self-reporting users in total, in those cases we used all matches.

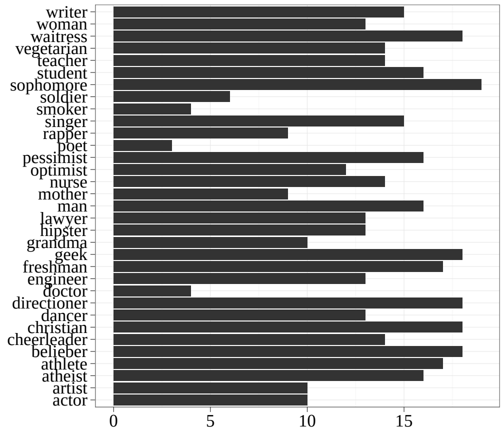

Figure 2: Valid self-identifying tweets from sample of 20.

is noteworthy that a non-trivial number for each were judged as actually self-identifying.

Indicative Language Most work in user classification relies on featurizing language use, most simply through binary indicators recording whether a user did or did not use a particular word in a history of n tweets. To explore whether language provides signal for future work in fine-grain social role prediction, we constructed a set of experiments, one per role, where training and test sets were balanced between users from a random background sample and self-reported users. Baseline accuracy in these experiments was thus 50%.

Each training set had a target of 600 users (300 background, 300 self-identified); for those roles with less than 300 users self-identifying, all users were used, with an equal number background. We used the `Jerboa` (Van Durme, 2012a) platform to convert data to binary feature vectors over a unigram vocabulary filtered such that the minimum frequency was 5 (across unique users). Training and testing was done with a log-linear model via `LibLinear` (Fan et al., 2008). We used the positively annotated data to form test sets, balanced with data from the background set. Each test set had a theoretical maximum size of 40, but for several classes it was in the single digits (see Figure 2). Despite the varied noisiness of our simple pattern-bootstrapped training data, and the small size of our annotated test set, we see in Figure 3 that we are able to successfully achieve statistically significant predictions of social role for the majority of our selected examples.

Table 3 highlights examples of language indicative of role, as determined by the most positively weighted unigrams in the classification experiment. These results qualitatively suggest many roles under consideration may be teased out from a background population by focussing on language that follows expected use patterns. For example the use of the term *game* by athletes, *studio* by artists, *mixtape* by rappers, or *jesus* by Christians.

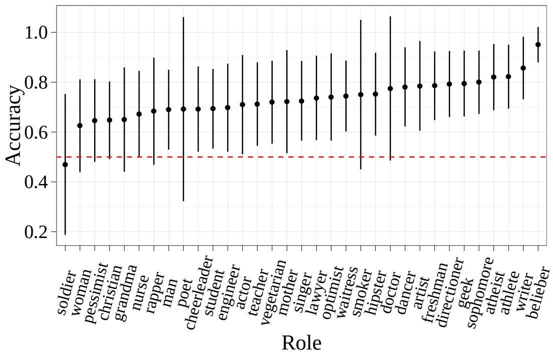

Figure 3: Accuracy in classifying social roles.

Role :: Feature (_Rank)
artist morning, summer, life, most, amp, studio
atheist fuck, fucking, shit, makes, dead, ..., religion$_{19}$
athlete lol, game, probably, life, into, ..., team$_{9}$
belieber justin, justinbeiber, believe, beliebers, bieber
cheerleader cheer, best, excited, hate, mom, ..., prom$_{16}$
christian lol, ..., god$_{12}$, pray$_{13}$, ..., bless$_{17}$, ..., jesus$_{20}$
dancer dance, since, hey, never, been
directioner harry, d, follow, direction, never, liam, niall
doctor sweet, oh, or, life, nothing
engineer (, then, since, may,), test$_{9}$, -$_{17}$, =$_{18}$
freshman summer, homework, na, ..., party$_{19}$, school$_{20}$
geek trying, oh, different, dead, been
grandma morning, baby, around, night, excited
hipster fucking, actually, thing, fuck, song
lawyer did, never, his, may, pretty, law, even, office
man man, away, ai, young, since
mother morning, take, fuck, fucking, trying
nurse lol, been, morning, ..., night$_{10}$, nursing$_{11}$, shift$_{13}$
optimist morning, enough, those, everything, never
poet feel, song, even, say, yo
rapper fuck, morning, lol, ..., mixtape$_{8}$, songs$_{15}$
singer sing, song, music, lol, never
smoker fuck, shit, fucking, since, ass, smoke, weed$_{20}$
solider ai, beautiful, lol, wan, trying
sophmore summer, >, ..., school$_{11}$, homework$_{12}$
student anything, summer, morning, since, actually
teacher teacher, morning, teach, ..., students$_{7}$, ..., school$_{20}$
vegetarian actually, dead, summer, oh, morning
waitress man, try, goes, hate, fat
woman lol, into, woman, morning, never
writer write, story, sweet, very, working

Table 3: Most-positively weighted features per role, along with select features within the top 20. Surprising **mother** features come from ambigious self-identification, as seen in tweets such as: *I'm a mother f!cking starrrrr.*

3 Characteristic Attributes

Bergsma and Van Durme (2013) showed that the task of mining attributes for conceptual classes can relate straightforwardly to author attribute prediction. If one views a role, in their case gender, as two conceptual classes, **male** and **female**, then existing attribute extraction methods for third-person content (e.g., news articles) can be cheaply used to create a set of bootstrapping features for building classifiers over first-person content (e.g., tweets). For example, if we learn from news corpora that: *a man may have a wife*, then a tweet saying: *...my wife...* can be taken as potential evidence of membership in the **male** conceptual class.

In our second study, we test whether this idea extends to our wider set of fine-grained roles. For example, we aimed to discover that a **doctor** may *have a patient*, while a **hairdresser** may *have a salon*; these properties can be expressed in first-person content as possessives like *my patient* or *my salon*. We approached this task by selecting target roles from the first experiment and ranking characteristic attributes for each using pointwise mutual information (PMI) (Church and Hanks, 1990).

First, we counted all terms matching a target social role's possessive pattern (e.g., *doctor's* __) in the web-scale n-gram corpus Google V2 (Lin et al., 2010)[5]. We ranked the collected terms by computing PMI between classes and attribute terms. Probabilities were estimated from counts of the class-attribute pairs along with counts matching the generic possessive patterns *his* __ and *her* __ which serve as general background categories. Following suggestions by Bergsma and Van Durme, we manually filtered the ranked list.[6] We removed attributes that were either (a) not nominal, or (b) not indicative of the social role. This left fewer than 30 attribute terms per role, with many roles having fewer than 10.

We next performed a precision test to identify potentially useful attributes in these lists. We examined tweets with a first person possessive pattern for each attribute term from a small corpus of tweets collected over a single month in 2013, discarding those attribute terms with no positive matches. This precision test is useful regardless of how attribute lists are generated. The attribute term *chart*, for example, had high PMI with **doctor**; but a precision test on the phrase *my chart* yielded a single tweet which referred not to a medical chart but to a top ten list (prompting removal of this attribute). Using this smaller high-precision set of attribute terms, we collected tweets from the Twitter Firehose over the period 2011-2013.

4 Attribute-based Classification

Attribute terms are less indicative overall than self-ID, e.g., the phrase *I'm a barber* is a clearer signal than *my scissors*. We therefore include a role verification step in curating a collection of positively identified users. We use the crowdsourcing platform Mechanical Turk[7] to judge whether the person tweeting held a given role Tweets were judged 5-way redundantly. Mechanical Turk judges ("Turkers") were presented with a tweet and the prompt: *Based on this tweet, would you think this person is a* BARBER/HAIRDRESSER*?* along with four response options: *Yes*, *Maybe*, *Hard to tell*, and *No*.

We piloted this labeling task on 10 tweets per attribute term over a variety of classes. Each answer was associated with a score (Yes = 1, Maybe = .5, Hard to tell = No = 0) and aggregated across the five judges. We found in development that an aggregate score of 4.0 (out of 5.0) led to an acceptable agreement rate between the Turkers and the experimenters, when the tweets were randomly sampled and judged internally. We found that making conceptual class assignments based on a single tweet was often a subtle task. The results of this labeling study are shown in Figure 4, which gives the percent of tweets per attribute that were 4.0 or above. Attribute terms shown in red were manually discarded as being inaccurate (low on the y-axis) or non-prevalent (small shape).

From the remaining attribute terms, we identified users with tweets scoring 4.0 or better as positive examples of the associated roles. Tweets from those users were scraped via the Twitter API to construct corpora for each role. These were split intro train and test, balanced with data from the same background set used in the self-ID study.

Test sets were usually of size 40 (20 positive, 20 background), with a few classes being sparse (the smallest had only 16 instances). Results are shown in Figure 5. Several classes in this balanced setup can be predicted with accuracies in the 70-90%

[5] In this corpus, follower-type roles like **belieber** and **directioner** are not at all prevalent. We therefore focused on occupational and habitual roles (e.g., **doctor, smoker**).

[6] Evidence from cognitive work on memory-dependent tasks suggests that such relevance based filtering (recognition) involves less cognitive effort than generating relevant attributes (recall) see (Jacoby et al., 1979). Indeed, this filtering step generally took less than a minute per class.

[7] https://www.mturk.com/mturk/

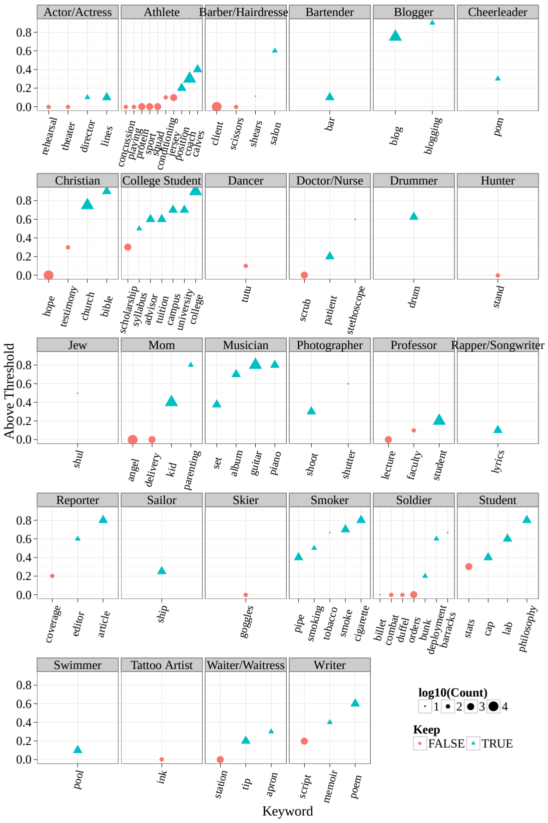

Figure 4: Turker judged quality of attributes selected as candidate features for bootstrapping positive instances of the given social role.

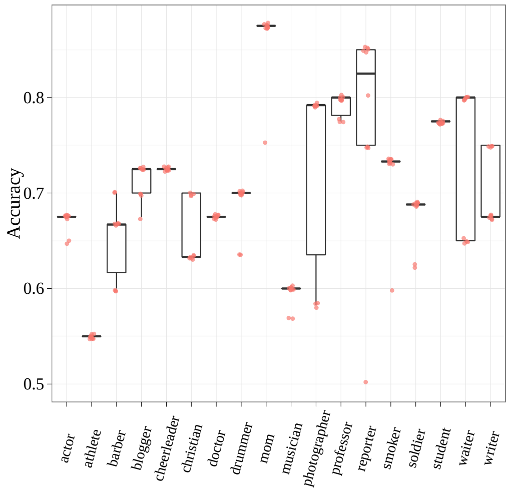

Figure 5: Classifier accuracy on balanced set contrasting agreed upon Twitter users of a given role against users pulled at random from the 1% stream.

range, supporting our claim that there is discriminating content for a variety of these social roles.

Conditional Classification How accurately we can predict membership in a given class when a Twitter user sends a tweet matching one of the targeted attributes? For example, if one sends a tweet saying *my coach*, then how likely is it that author is an **athlete**?

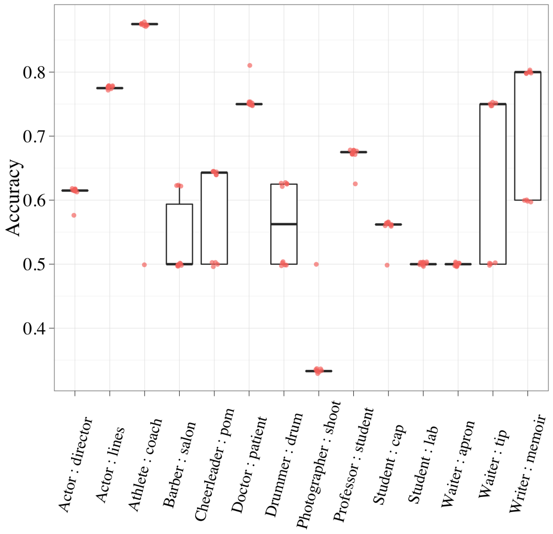

Figure 6: Results of positive vs negative by attribute term. Given that a user tweets ... *my lines* ... we are nearly 80% accurate in identifying whether or not the user is an actor.

Using the same collection as the previous experiment, we trained classifiers conditioned on a given attribute term. Positive instances were taken to be those with a score of 4.0 or higher, with negative instances taken to be those with scores of 1.0 or lower (strong agreement by judges that the original tweet did not provide evidence of the given role). Classification results are shown in Figure 6.

5 Conclusion

We have shown that Twitter contains sufficiently robust signal to support more fine-grained author attribute prediction tasks than have previously been attempted. Our results are based on simple, intuitive search patterns with minimal additional filtering: this establishes the feasibility of the task, but leaves wide room for future work, both in the sophistication in methodology as well as the diversity of roles to be targeted. We exploited two complementary types of indicators: *self-identification* and *self-possession* of conceptual class (role) attributes. Those interested in identifying latent demographics can extend and improve these indicators in developing ways to identify groups of interest within the general population of Twitter users.

Acknowledgements This material is partially based on research sponsored by the NSF under grants DGE-123285 and IIS-1249516 and by DARPA under agreement number FA8750-13-2-0017 (the DEFT program).

References

Abdulrahman Almuhareb and Massimo Poesio. 2004. Attribute-based and value-based clustering: an evaluation. In *Proceedings of EMNLP*.

Shane Bergsma and Benjamin Van Durme. 2013. Using Conceptual Class Attributes to Characterize Social Media Users. In *Proceedings of ACL*.

Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clay Fink, and Theresa Wilson. 2012. Language identification for creating language-specific twitter collections. In *Proceedings of the NAACL Workshop on Language and Social Media*.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of EMNLP*.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. In *ICWSM*.

Jacob Eisenstein, Brendan O'Connor, Noah Smith, and Eric P. Xing. 2010. A latent variable model of geographical lexical variation. In *Proceedings of EMNLP*.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsief, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, (9).

Nikesh Garera and David Yarowsky. 2009. Modeling latent biographic attributes in conversational genres. In *Proceedings of ACL*.

Larry L Jacoby, Fergus IM Craik, and Ian Begg. 1979. Effects of decision difficulty on recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, 18(5):585–600.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*.

Alex Lamb, Michael J. Paul, and Mark Dredze. 2013. Separating fact from fear: Tracking flu infections on twitter. In *Proceedings of NAACL*.

Dekang Lin, Kenneth Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil, Emily Pitler, Rachel Lathbury, Vikram Rao, Kapil Dalwani, and Sushant Narsale. 2010. New tools for web-scale n-grams. In *Proc. LREC*, pages 2221–2227.

Saif M. Mohammad, Svetlana Kiritchenko, and Joel Martin. 2013. Identifying purpose behind electoral tweets. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '13, pages 1–9.

Dong Nguyen, Noah A Smith, and Carolyn P Rosé. 2011. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 115–123. Association for Computational Linguistics.

Marius Paşca and Benjamin Van Durme. 2008. Weakly-Supervised Acquisition of Open-Domain Classes and Class Attributes from Web Documents and Query Logs. In *Proceedings of ACL*.

Michael J Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *ICWSM*.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. Democrats, Republicans and Starbucks afficionados: User classification in Twitter. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 430–438. ACM.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the Workshop on Search and Mining User-generated Contents (SMUC)*.

Lenhart K. Schubert. 2002. Can we derive general world knowledge from texts? In *Proceedings of HLT*.

Benjamin Van Durme. 2012a. Jerboa: A toolkit for randomized and streaming algorithms. Technical Report 7, Human Language Technology Center of Excellence, Johns Hopkins University.

Benjamin Van Durme. 2012b. Streaming analysis of discourse participants. In *Proceedings of EMNLP*.

Jennifer Wortman. 2008. Viral marketing and the diffusion of trends on social networks. Technical Report MS-CIS-08-19, University of Pennsylvania, May.

Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. In *Proceedings of ICWSM*.