

# Predicting Fine-grained Social Roles with Selectional Preferences

**Charley Beller**      **Craig Harman**      **Benjamin Van Durme**  
charleybeller@jhu.edu    craig@craigharman.net    vandurme@cs.jhu.edu  
Human Language Technology Center of Excellence  
Johns Hopkins University, Baltimore, MD USA

## Abstract

Selectional preferences, the tendencies of predicates to select for certain semantic classes of arguments, have been successfully applied to a number of tasks in computational linguistics including word sense disambiguation, semantic role labeling, relation extraction, and textual inference. Here we leverage the information encoded in selectional preferences to the task of predicting fine-grained categories of authors on the social media platform Twitter. First person uses of verbs that select for a given social role as subject (e.g. *I teach ...* for **teacher**) are used to quickly build up binary classifiers for that role.

## 1 Introduction

It has long been recognized that linguistic predicates preferentially select arguments that meet certain semantic criteria (Katz and Fodor, 1963; Chomsky, 1965). The verb *eat* for example selects for an animate subject and a comestible object. While the information encoded by selectional preferences can and has been used to support natural language processing tasks such as word sense disambiguation (Resnik, 1997), syntactic disambiguation (Li and Abe, 1998) and semantic role labeling (Gildea and Jurafsky, 2002), much of the work on the topic revolves around developing methods to induce selectional preferences from data. In this setting, end-tasks can be used for evaluation of the resulting collection. Ritter et al. (2010) gave a recent overview of this work, breaking it down into class-based approaches (Resnik, 1996; Li and Abe, 1998; Clark and Weir, 2002; Pantel et al., 2007), similarity-based approaches (Dagan et al., 1999; Erk, 2007), and approaches using discriminative (Bergsma et al., 2008) or generative probabilistic models (Rooth et al., 1999) like their own.

One of our contributions here is to show that the literature on selectional preferences relates to the analysis of the first person content transmitted through social media. We make use of a “quick and dirty” method for inducing selectional preferences and apply the resulting collections to the task of predicting fine-grained latent *author attributes* on Twitter. Our method for inducing selectional preferences is most similar to class-based approaches, though unlike approaches such as by Resnik (1996) we do not require a WordNet-like ontology.

The vast quantity of informal, first-person text data made available by the rise of social media platforms has encouraged researchers to develop models that predict broad user categories like age, gender, and political preference (Garera and Yarowsky, 2009; Rao et al., 2010; Burger et al., 2011; Van Durme, 2012b; Zamal et al., 2012). Such information is useful for large scale demographic research that can fuel computational social science advertising.

Similarly to Beller et al. (2014), we are interested in classification that is finer-grained than gender or political affiliation, seeking instead to predict *social roles* like *smoker*, *student*, and *artist*. We make use of a light-weight, unsupervised method to identify selectional preferences and use the resulting information to rapidly bootstrap classification models.

## 2 Inducing Selectional Preferences

Consider the task of predicting social roles in more detail: For a given role, e.g. **artist**, we want a way to distinguish role-bearing from non-role-bearing users. We can view each social role as being a fine-grained version of a semantic class of the sort required by class-based approaches to selectional preferences (e.g. the work by Resnik (1996) and those reviewed by Light and Greiff (2002)). The goal then is to identify a set of verbs that preferen-

tially select that particular class as argument. Once we have a set of verbs for a given role, simple pattern matches against first person subject templates like *I \_\_\_* can be used to identify authors that bear that social role.

In order to identify verbs that select for a given role  $r$  as subject we use an unsupervised method inspired by Bergsma and Van Durme (2013) that extracts features from third-person content (i.e. newswire) to build classifiers on first-person content (i.e. tweets). For example, if we read in a news article that *an artist drew ...*, we can take a tweet saying *I drew ...* as potential evidence that the author bears the **artist** social role.

We first count all verbs  $v$  that appear with role  $r$  as subject in the web-scale, part-of-speech tagged n-gram corpus, Google V2 (Lin et al., 2010). The resulting collection of verbs is then ranked by computing their pointwise mutual information (Church and Hanks, 1990) with the subject role  $r$ . The PMI of a given role  $r$  and a verb  $v$  that takes  $r$  as subject is given as:

$$\text{PMI}(r, v) = \log \frac{P(r, v)}{P(r)P(v)}$$

Probabilities are estimated from counts of the role-verb pairs along with counts matching the generic subject patterns *he \_\_\_* and *she \_\_\_* which serve as general background cases. This gives us a set of verbs that preferentially select for the subset of persons filling the given role.

The output of the PMI ranking is a high-recall list of verbs that preferentially select the given social role as subject over a background population. Each such list then underwent a manual filtering step to rapidly remove non-discriminative verbs and corpus artifacts. One such artifact from our corpus was the term *wannabe* which was spuriously elevated in the PMI ranking based on the relative frequency of the bigram *artist wannabe* as compared to *she wannabe*. Note that in the first case *wannabe* is best analyzed as a noun, while in the second case a verbal analysis is more plausible. The filtering was performed by one of the authors and generally took less than two minutes per list. The rapidity of the filtering step is in line with findings such as by Jacoby et al. (1979) that relevance based filtering involves less cognitive effort than generation. After filtering the lists contained fewer than 40 verbs selecting each social role.

In part because of the pivot from third- to first-person text we performed a precision test on the

remaining verbs to identify which of them are likely to be useful in classifying twitter users. For each remaining verb we extracted all tweets that contained the first person subject pattern *I \_\_\_* from a small corpus of tweets drawn from the free public 1% sample of the Twitter Firehose over a single month in 2013. Verbs that had no matches which appeared to be composed by a member of the associated social role were discarded. Using this smaller high-precision set of verbs, we collected tweets from a much larger corpus drawn from 1% sample over the period 2011-2013.

One notable feature of the written English in social media is that sentence subjects can be optionally omitted. Subject-drop is a recognized feature of other informal spoken and written registers of English, particularly ‘diary dialects’ (Thrasher, 1977; Napoli, 1982; Haegeman and Ihsane, 2001; Weir, 2012; Haegeman, 2013; Scott, 2013). Because of the prevalence of subjectless cases we collected two sets of tweets: those matching the first person subject pattern *I \_\_\_* and those where the verb was tweet initial. Example tweets for each of our social roles can be seen in Table 2.

### 3 Classification via selectional preferences

We conducted a set of experiments to gauge the strength of the selectional preference indicators for each social role. For each experiment we used balanced datasets for training and testing with half of the users taken from a random background sample and half from a collection of users identified as belonging to the social role. Base accuracy was thus 50%.

To curate the collections of positively identified users we crowdsourced a manual verification procedure. We use the popular crowdsourcing platform Mechanical Turk<sup>1</sup> to judge whether, for a tweet containing a given verb, the author held the role that verb prefers as subject. Each tweet was judged using 5-way redundancy.

Mechanical Turk judges (“Turkers”) were presented with a tweet and the prompt: *Based on this tweet, would you think this person is a ARTIST?* along with four response options: *Yes*, *Maybe*, *Hard to tell*, and *No*. An example is shown in Figure 1.

We piloted this labeling task with a goal of 20 tweets per verb over a variety of social roles.

<sup>1</sup><https://www.mturk.com/mturk/>

<b>Artist</b> <i>draw</i>	Yeaa this a be the first time I draw my shit onn
<b>Athlete</b> <i>play</i>	@[user] @[user] i have got the night off tonight because I played last night and I am going out for dinner so won't be able to come"
<b>Blogger</b> <i>blogged</i>	@[user] I decided not to renew. I blogged about it on the fan club. a bit shocked no neg comments back to me
<b>Cheerleader</b> <i>cheer</i>	I really dont wanna cheer for this game I have soo much to do
<b>Christian</b> <i>thank</i>	Had my bday yesterday 3011 nd had a good night with my friends. I thank God 4 His blessings in my life nd praise Him 4 adding another year.
<b>DJ</b> <i>spin</i>	Quick cut session before I spin tonight
<b>Filmmaker</b> <i>film</i>	@[user] apparently there was no audio on the volleyball game I filmed so...there will be no "NAT sound" cause I have no audio at all
<b>Media Host</b> <i>interview</i>	Oh. I interviewed her on the @[user] . You should listen to the interview. Its awesome! @[user] @[user] @[user]
<b>Performer</b> <i>perform</i>	I perform the flute... kareem shocked...
<b>Producer</b> <i>produce</i>	RT @[user]: Wow 2 films in Urban-world this year-1 I produced ... [URL]
<b>Smoker</b> <i>smoke</i>	I smoke , i drank .. was my shit bra !
<b>Stoner</b> <i>puff</i>	I'm a cigarello fiend smokin weed like its oxygen Puff pass, nigga I puff grass till I pass out
<b>Student</b> <i>finish</i>	I finish school in March and my friend birthday in March ...
<b>Teacher</b> <i>teach</i>	@[user] home schooled I really wanna find out wat it's like n making new friends but home schooling is cool I teach myself mums ill

Table 1: Example verbs and sample tweets collected using them in the first person subject pattern (*I ...*).

Each answer was associated with a score (Yes = 1, Maybe = .5, Hard to tell = No = 0) and aggregated across the five judges, leading to a range of possible scores from 0.0 to 5.0 per tweet. We found in development that an aggregate score of 4.0 led to an acceptable agreement rate between the Turkers and the experimenters, when the tweets were randomly sampled and judged internally.

Verbs were discarded for being either insufficiently accurate or insufficiently prevalent in the corpus. From the remaining verbs, we identified users with tweets scoring 4.0 or better as the positive examples of the associated social roles. These positively identified user’s tweets were scraped using the Twitter API in order to construct user-specific corpora of positive examples for each role.

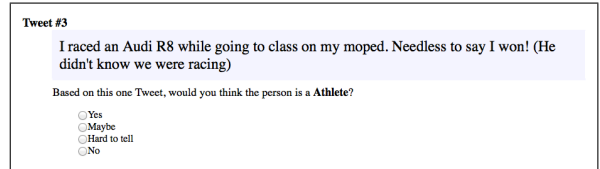


Figure 1: Mechanical Turk presentation

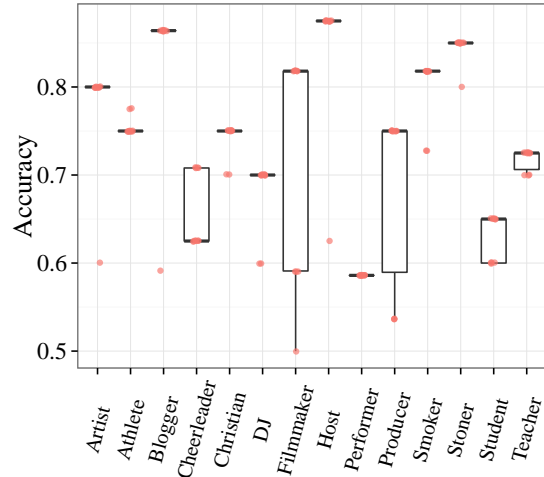


Figure 2: Accuracy of classifier trained and tested on balanced set contrasting agreed upon Twitter users of a given role, against users pulled at random from the 1% stream.

### 3.1 General Classification

The positively annotated examples were balanced with data from a background set of Twitter users to produce training and test sets. These test sets were usually of size 40 (20 positive, 20 background), with a few classes being sparser (the smallest test set had only 28 instances). We used the `Jerboa` (Van Durme, 2012a) platform to convert our 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). Results are shown in Figure 2. As can be seen, a variety of classes in this balanced setup can be predicted with accuracies in the range of 80%. This shows that the information encoded in selectional preferences contains discriminating signal for a variety of these social roles.

### 3.2 Conditional Classification

How accurately can we predict membership in a given class when a Twitter user sends a tweet matching one of the collected verbs? For example, if one sends a tweet saying *I race ...*, then how likely is it that the author is an **athlete**?

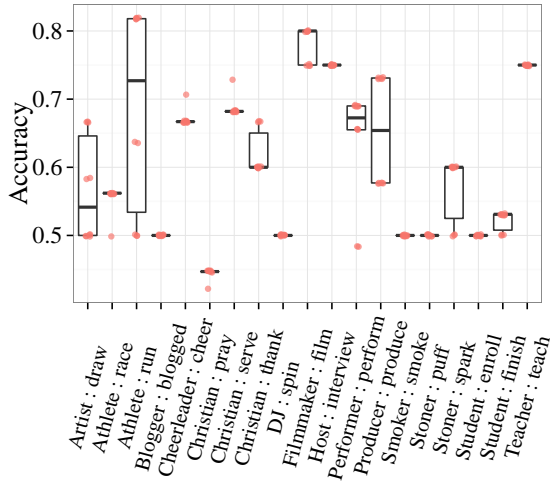


Figure 3: Results of positive vs negative by verb. Given that a user writes a tweet containing *I interview ...* or *Interviewing ...* we are about 75% accurate in identifying whether or not the user is a Radio/Podcast Host.

# Users	# labeled	# Pos	# Neg	Attribute
199022	516	63	238	Artist-draw
45162	566	40	284	Athlete-race
1074289	1000	54	731	Athlete-run
9960	15	14	0	Blogger-blog
2204	140	57	18	College Student-enroll
247231	1000	85	564	College Student-finish
60486	845	61	524	Cheerleader-cheer
448738	561	133	95	Christian-pray
92223	286	59	180	Christian-serve
428337	307	78	135	Christian-thank
17408	246	17	151	DJ-spin
153793	621	53	332	Filmmaker-film
36991	554	42	223	Radio Host-interview
43997	297	81	97	Performer-perform
69463	315	71	100	Producer-produce
513096	144	74	8	Smoker-smoke
5542	124	49	15	Stoner-puff
5526	229	59	51	Stoner-spark
149244	495	133	208	Teacher-teach

Table 2: Numbers of positively and negatively identified users by indicative verb.

Using the same collection as the previous experiment, we trained classifiers conditioned on a given verb 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 3. Note that for a number of verb terms these thresholds left very sparse collections of users. There were only 8 users, for example, that tweeted the phrase *I smoke ...* but were labeled as negative instances of Smokers. Counts are given in Table 2.

Despite the sparsity of some of these classes, many of the features learned by our classifiers make intuitive sense. Highlights of the most highly weighted unigrams from the classification

Verb	Feature (.Rank)
draw	drawing, art, book <sub>4</sub> , sketch <sub>14</sub> , paper <sub>19</sub>
race	race, hard, winter, won <sub>11</sub> , training <sub>16</sub> , run <sub>17</sub>
run	awesome, nike <sub>6</sub> , fast <sub>9</sub> , marathon <sub>20</sub>
blog	notes, boom, hacked <sub>4</sub> , perspective <sub>9</sub>
cheer	cheer, pictures, omg, text, literally
pray	through, jesus <sub>3</sub> , prayers <sub>7</sub> , lord <sub>14</sub> , thank <sub>17</sub>
serve	lord, jesus, church, blessed, pray, grace
thank	[ ], blessed, lord, trust <sub>11</sub> , pray <sub>12</sub>
enroll	fall, fat, carry, job, spend, fail <sub>15</sub>
finish	hey, wrong, may <sub>8</sub> , move <sub>9</sub> , officially <sub>14</sub>
spin	show, dj, music, dude, ladies, posted, listen
film	please, wow, youtube, send, music <sub>8</sub>
perform	[ ], stuck, act, song, tickets <sub>7</sub> , support <sub>16</sub>
produce	follow, videos, listen <sub>10</sub> , single <sub>11</sub> , studio <sub>13</sub> ,
interview	fan, latest, awesome, seems
smoke	weakness, runs, ti, simply
puff	bout, \$ <sub>7</sub> , smh <sub>9</sub> , weed <sub>10</sub>
spark	dont, fat <sub>5</sub> , blunt <sub>6</sub> , smoke <sub>11</sub>
teach	forward, amazing, students, great, teacher <sub>7</sub>

Table 3: Most-highly indicative features that a user holds the associated role given that they used the phrase *I VERB* along with select features within the top 20.

experiments are shown in Table 3. Taken together these features suggest that several of our roles can be distinguished from the background population by focussing on typical language use. The use of terms like, e.g., *sketch* by artists, *training* by athletes, *jesus* by Christians, and *students* by teachers conforms to expected pattern of language use.

## 4 Conclusion

We have shown that verb-argument selectional preferences relates to the content-based classification strategy for latent author attributes. In particular, we have presented initial studies showing that mining selectional preferences from third-person content, such as newswire, can be used to inform latent author attribute prediction based on first-person content, such as that appearing in social media services like Twitter.

Future work should consider the question of *priors*. Our study here relied on balanced class experiments, but the more fine-grained the social role, the smaller the subset of the population we might expect will possess that role. Estimating these priors is thus an important point for future work, especially if we wish to couple such demographic predictions within a larger automatic system, such as the aggregate prediction of targeted sentiment (Jiang et al., 2011).

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

## References

- Charley Beller, Rebecca Knowles, Craig Harman, Shane Bergsma, Margaret Mitchell, and Benjamin Van Durme. 2014. I'm a believer: Social roles via self-identification and conceptual attributes. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.
- Shane Bergsma and Benjamin Van Durme. 2013. Using Conceptual Class Attributes to Characterize Social Media Users. In *Proceedings of ACL*.
- Shane Bergsma, Dekang Lin, and Randy Goebel. 2008. Discriminative learning of selectional preference from unlabeled text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 59–68. Association for Computational Linguistics.
- John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *Proceedings of EMNLP*.
- Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. Number 11. MIT press.
- Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.
- Stephen Clark and David Weir. 2002. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2).
- Ido Dagan, Lillian Lee, and Fernando CN Pereira. 1999. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1-3):43–69.
- Katrin Erk. 2007. A simple, similarity-based model for selectional preferences. In *Proceeding of the 45th Annual Meeting of the Association for Computational Linguistics*, volume 45, page 216.
- Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, 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*.
- Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.
- Liliane Haegeman and Tabea Ihsane. 2001. Adult null subjects in the non-pro-drop languages: Two diary dialects. *Language acquisition*, 9(4):329–346.
- Liliane Haegeman. 2013. The syntax of registers: Diary subject omission and the privilege of the root. *Lingua*, 130:88–110.
- 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.
- Long Jiang, Mo Yu, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of ACL*.
- Jerrold J Katz and Jerry A Fodor. 1963. The structure of a semantic theory. *language*, pages 170–210.
- Hang Li and Naoki Abe. 1998. Generalizing case frames using a thesaurus and the MDL principle. *Computational linguistics*, 24(2):217–244.
- Marc Light and Warren Greiff. 2002. Statistical models for the induction and use of selectional preferences. *Cognitive Science*, 26(3):269–281.
- 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.
- Donna Jo Napoli. 1982. Initial material deletion in English. *Glossa*, 16(1):5–111.
- Patrick Pantel, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski, and Eduard H Hovy. 2007. ISP: Learning inferential selectional preferences. In *HLT-NAACL*, pages 564–571.
- 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)*.
- Philip Resnik. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61(1):127–159.
- Philip Resnik. 1997. Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*, pages 52–57. Washington, DC.
- Alan Ritter, Masaum, and Oren Etzioni. 2010. A latent dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 424–434. Association for Computational Linguistics.
- Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pages 104–111. Association for Computational Linguistics.

- Kate Scott. 2013. Pragmatically motivated null subjects in English: A relevance theory perspective. *Journal of Pragmatics*, 53:68–83.
- Randolph Thrasher. 1977. One way to say more by saying less: A study of so-called subjectless sentences. *Kwansei Gakuin University Monograph Series*, 11.
- 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*.
- Andrew Weir. 2012. Left-edge deletion in English and subject omission in diaries. *English Language and Linguistics*, 16(01):105–129.
- 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*.