What's in a Name? Using First Names as Features for Gender Inference in Twitter

Wendy Liu and **Derek Ruths**
School of Computer Science
McGill University
wendy.liu@mail.mcgill.ca and *derek.ruths@mcgill.ca*

Abstract

Despite significant work on the problem of inferring a Twitter user's gender from her online content, no systematic investigation has been made into leveraging the most obvious signal of a user's gender: first name. In this paper, we perform a thorough investigation of the link between gender and first name in English tweets. Our work makes several important contributions.

The first and most central contribution is two different strategies for incorporating the user's self-reported name into a gender classifier. We find that this yields a 20% increase in accuracy over a standard baseline classifier. These classifiers are the most accurate gender inference methods for Twitter data developed to date.

In order to evaluate our classifiers, we developed a novel way of obtaining gender-labels for Twitter users that does not require analysis of the user's profile or textual content. This is our second contribution. Our approach eliminates the troubling issue of a label being somehow derived from the same text that a classifier will use to infer the label.

Finally, we built a large dataset of gender-labeled Twitter users and, crucially, have published this dataset for community use. To our knowledge, this is the first gender-labeled Twitter dataset available for researchers. Our hope is that this will provide a basis for comparison of gender inference methods.

Introduction

Many people in industry and academia share an intense interest in being able to obtain the demographic breakdown of a population of Twitter users. Such information will offer social scientists new perspective into the online communities they study and provide consumer-based companies with more actionable features of their current and potential customer base. Few demographic variables are more fundamental and discrete than gender. As a result, the inference of a Twitter user's gender has become an active area of research (e.g., (Burger, Henderson, and Zarrella 2011; Liu, Zamal, and Ruths 2012; Zamal, Liu, and Ruths 2012; Pennacchiotti and Popescu 2011)).

The majority of recent work on gender inference has focused on improving the classification machinery rather than on expanding or refining the features that a classifier uses to infer a gender label. Commonly-used features primarily involve statistics of the user-generated text including the frequency of n-gram subsamples as well as the most frequent words, hashtags, and mentions of other users (Burger, Henderson, and Zarrella 2011; Liu, Zamal, and Ruths 2012; Zamal, Liu, and Ruths 2012; Pennacchiotti and Popescu 2011). Most notably omitted in most methods is any kind of encoding of the Twitter user's self-reported name. In fact, to our knowledge, only one existing method has even approached the task of incorporating this feature in some way (Burger, Henderson, and Zarrella 2011).

From a purely anecdotal perspective, we observe that first names can be a strong signal of an individual's gender: for example, diminishingly few girls are named "Derek" and boys named "Wendy" are exceedingly hard to find. As we show later, census data confirms and quantifies this intuition across a wide array of names. Given this fact, we should expect that gender inference methods will benefit from incorporating the user's first name in the user's feature set. Thus, it is quite surprising that only one existing method has attempted to incorporate this feature. This lack of systematic investigation into the use of the user's name in gender inference provided the motivation for the present study.

In this paper, our primary goal is to characterize how incorporating a user's name into a gender classifier improves the quality of inferred labels. In order to do this, we designed and implemented two gender-inference methods which leverage a user's name in different ways. Both reduce the individual's name to its observed gender association (e.g. 1 for a name that is always given to males, -1 if always given to females, and 0 if given with equal likelihood to both genders). In one method, the name association score was included as an element of the user's feature vector. In the other, the score was used as a high-level switch to determine whether to use a feature-based classifier at all. The intuition behind the latter method is that a user who reports the name "Bob" is so strongly signaling the male gender that processing the feature vector is, statistically speaking, unnecessary.

In order to compare the performance of these two strategies against one another and against a baseline method (which ignored name information entirely), we built a

gender-labeled Twitter dataset. Unlike past methods which have depended on the user's generated textual content in one form or another (e.g., (Burger, Henderson, and Zarrella 2011; Zamal, Liu, and Ruths 2012; Pennacchiotti and Popescu 2011)), our novel approach uses the account profile picture as a signal for user gender. Amazon Mechanical Turk workers were used to identify the gender of the individual depicted in the profile picture. We consider this method and the dataset we assembled (and have now made public) to be two ancillary, yet important, contributions of this work.

When run on the assembled dataset, we found that including a user's first name significantly improved the inferred accuracy: from a baseline of 83% to 87% reported by our new methods. This constitutes a 20% increase in the accuracy of inferred gender labels as well as a significant improvement in the performance reported by other methods (81%).

From a broader perspective, the present investigation has implications for gender inference on many other online platforms besides Twitter. First, we have devised a novel method by which Amazon Mechanical Turk may be used to obtain high quality gender-labeled data from user profile data. Second, we have shown that a user's first name is a valuable feature when inferring gender labels. Furthermore, where the user's name is concerned, our methods are agnostic to the platform being analyzed. Since a first name (or nickname) is a commonly supported field in a user profiles on online social platforms and discussion forums, the methods and results discussed here are directly applicable to the gender inference problem on a wide array of online environments.

Prior work

Gender inference on Twitter Much work has been done on the problem of gender inference on Twitter. Without exception, the top performing methods in this area use feature-based classifiers such as support vector machines and boosted decision trees (Burger, Henderson, and Zarrella 2011; Zamal, Liu, and Ruths 2012; Pennacchiotti and Popescu 2011). Probabilistic models such as Naive Bayes and latent semantic analysis have also been considered, but have been used primarily as features themselves that are fed to the former set of classifiers.

Research in this area has primarily focused on the construction of more sophisticated classification systems; this is to say that the elements composing user feature vectors are fairly static across papers. Standard features include n-grams contained in user tweets, most frequently used words, hashtags, and mentioned users, and basic statistics concerning user following habits, popularity, and tweeting frequency (e.g., (Burger, Henderson, and Zarrella 2011; Zamal, Liu, and Ruths 2012)). Thus, in general, we see a significant opportunity for devising new, useful features for inclusion in the classification process (for gender, as well as for other features).

Most relevant to the present study is the work of Burger et al. which investigates the contribution of a variety of features to the gender inference problem. Among these features are a set (of n-grams) derived from the user's full name. Using the winnow algorithm, they obtain the remarkable finding that using only n-grams extracted from the user's full name, their classifier yields a label assignment accuracy of 89% (Littlestone 1988). While the work is methodologically valid, it is important to recognize that the dataset used is not representative of the general Twitter population: they construct their dataset to include only users who (a) maintain a blog and (b) declare their gender in the profile of the blog. The authors themselves admit that this construction procedure likely creates a population that is not representative of general Twitter users; later in this paper, we prove this assertion by analyzing a reconstructed version of their dataset. In addition, we made several unsuccessful attempts to obtain the complete or partial dataset they used, making it virtually impossible for us to determine the extent to which their findings depend on the unique dataset creation methods employed. These facts, unfortunately, put us in a situation where we cannot presume that the results reported in their paper extend to the general Twitter population.

Amazon Mechanical Turk Amazon Mechanical Turk is a platform developed by Amazon for the distribution of short, simple tasks to legions of human workers around the world. A small payment is made (typically between \$0.01 and \$0.10 per task) on the completion of a *human intelligence task* (called a *HIT*). A number of studies have confirmed that Amazon Mechanical Turk can be a reliable way of obtaining high-quality annotations, labeling data, and even conducting more ambitious psychological and sociological experiments (Schnoebelen and Kuperman 2010; Buhrmester, Kwang, and Gosling 2011). In this regard, our work belongs to a growing set of studies that generate ground-truth labels for data using human annotations obtained through AMT.

Building the gender-labeled dataset

Because there are no canonical gender-labeled Twitter datasets available to the research community, we developed ours to use in this study. This is a universal practice among gender inference research on Twitter (e.g. (Burger, Henderson, and Zarrella 2011; Zamal, Liu, and Ruths 2012; Pennacchiotti and Popescu 2011)). We approached this dataset construction exercise, however, with several goals in mind beyond simply identifying a set of users to whom we could assign high confidence gender labels.

Selecting a representative sample of users In order to claim that a method's demonstrated performance generalizes to arbitrary Twitter users, one must show that the dataset used to evaluate the performance is representative. This is an important detail that is often overlooked (or at the very least not mentioned) by research in this area. One of our aims in this paper was to address this issue by showing that statistical features of our sampled dataset maintain the statistical features of Twitter at large.

Eliminating any correlation between labels and user content Most (and often all) features used by a latent

attribute classifier are derived from the text that a user generates either in her name, profile, or tweet history. Thus if labels are determined by anything contained in the user's textual content, there is the real possibility for the introduction of artificial linkage between the label and the textual content. Presumably such linkage will ultimately benefit the classifier and lead to inflation in the reported accuracy of the method. Studies which use manual annotation based on the user's content (e.g., (Conover et al. 2011)) or automated analysis of text generated by the user (e.g., (Burger, Henderson, and Zarrella 2011)) may be subject to this effect. We sought a data collection strategy that eliminates, as completely as possible, any dependency between a user's label assignment and her textual content. As will be discussed later in this section, we achieved this by deriving a user's gender label from her profile picture, which has no direct or systemic relationship to the profile and tweet text that she generated.

Releasing a public gender-labeled Twitter dataset While much work has been done on the gender inference problem on Twitter, there are no canonical datasets available to the community. As a result, every methodological paper develops its own dataset on which to evaluate the performance of its new approach. In general, dataset collection strategies differ across papers; and even when the strategies are similar, the exact user sets are different. This naturally makes the results of different studies rather incomparable. Certainly some qualitative assessments are possible, but ultimately a careful side-by-side comparison is impossible. In the interest of creating standardized datasets for the evaluation of gender inference methods, we have released the dataset that was produced for this study[1].

Gender-name association scores

Before discussing the dataset construction method, it is necessary to describe the gender-name association scores that were computed for a wide array of first names. These scores operationalize the intuitive notion that some names are very specific to a particular gender (e.g., "Weston", "Leena", and "Sonya") while others are common among males and females (e.g., "Kerry", "Tommie", and "Kris").

The gender association of name x is given by the formula $\frac{M(x)-F(x)}{M(x)+F(x)}$, where $M(x)$ is how many times the name is given to a male and $F(x)$ is the number of times the name has been given to a female. The score ranges from -1 (the name is only given to females) to 1 (the name is only given to males). One useful property of this definition of gender-name association score is that it is clear how to assign a score to a name for which we have no observations at all: such a name is given a score of 0, indicating that, a priori, there is equal likelihood that the name is associated with a female or male. Of course, it is possible that this a priori function could be improved using certain heuristics (e.g., names that end with an "a" tend to be associated with females). We identify such improvements as directions for future work.

In this project, we obtained $M(x)$ and $F(x)$ for a range of names x from name distributions collected from the 1990 US census[2]. Using names from the US census implicitly biases the coverage of names toward Western, American names. It would be possible to overcome this in future work by either incorporating name distributions from other countries or by learning the gender-name associations directly from the AMT-labeled data described next.

Dataset construction method

As has been alluded to earlier, a user's gender label was inferred based on his or her Twitter profile picture. In order to accelerate and standardize the process, this manual coding process was broken up into Amazon Mechanical Turk HITs. Each HIT consisted of 20 individual profile pictures that needed to be coded. For each profile picture in the HIT, the AMT worker indicated if they believed the profile picture depicted a male or a female. As many profile pictures do not convey the gender of the user (e.g., they are abstract graphics, pictures of celebrities, etc...), the worker also had the option of indicating that the gender of the user was unknown. Each profile picture was coded by 3 different AMT workers. A gender labeling was accepted only if all 3 AMT workers agreed on the same gender assignment.

Dataset construction This process was applied to $50,000$ Twitter users selected at random out of the Twitter gardenhose (the only condition applied was the requirement that the users have generated at least $1,000$ tweets over the lifespan of their Twitter account). Of these $50,000$ users, $12,681$ could be confidently assigned a gender label based on the AMT coding exercise. Given that our sample was random, we may estimate that 25.4% of all Twitter users have high-confidence, gender-relevant profile pictures.

Of the $12,681$ users assigned a gender label, $4,449$ were males and $8,232$ were females. Thus, only 35% of the users identified were males. This is a dramatically smaller number than the oft-cited estimate that 45% of Twitter users are male (Mislove et al. 2011). We have identified two possible explanations for this difference: it is certainly possible that the percent makeup of Twitter is somewhat different from the 45% estimate (although a 10% shift seems unlikely); it may be that female users are more likely to post informative profile pictures. Most likely the actual reason involves a combination of these two reasons - though a complete investigation into this would be a useful direction for future work. For the remainder of this study, wherever we use this dataset, we subsample the female population to create equal-sized male and female populations ($4,000$ users per gender).

In order to validate the quality of the AMT-generated labels, we looked at a subsample of the users whose self-reported names had 100% gender-name associations. For only this subsample, we computed the agreement

[1] Download at http://www.networkdynamics.org/...static/datasets/LiuRuthsMicrotext.zip

[2] http://www.census.gov/genealogy/www/data/...1990surnames/names_files.html

Table 1: The attributes of a random set of $100,000$ English tweets to those of the dataset constructed for this study and the dataset constructed in (Burger, Henderson, and Zarrella 2011). A comparison reveals that the Burger et al. dataset is a non-representative sample of Twitter users (on average they tweet significantly more and have a different distribution of name frequencies). In contrast, the dataset constructed using profile pictures models the features of random Twitter users more accurately, particularly where name frequencies are concerned.

Dataset	*Avg # tweets*	*% gender specific names*	*% census matched names*
Random	$13,221$	34.1%	37.3%
LiuRuths	$15,948$	33.8%	37.1%
Burger	$18,385$	**27.5%**	**29.4%**

between the AMT-assigned label and the gender associated with the user's name. For both males and females, disagreements occurred in less than 1% of all cases, suggesting that these labels have very high-confidence.

Following the labeling of the users, we collected the most recent $1,000$ tweets generated by each user. These tweets combined with the user's profile formed the textual content for the user.

Datasets as a representative sample of Twitter

In order to determine whether our dataset construction method obtains a representative sample of Twitter, we compared several average statistics over the dataset we constructed to the same statistics calculated over 100,000 randomly selected English users. As a point of reference, we also calculated these statistics for a dataset that was constructed similarly to that in (Burger, Henderson, and Zarrella 2011): since the original dataset was not shared, we obtained 10,000 English user accounts that linked to Wordpress and Blogger blogs (currently two of the most common blogging platforms). This was taken to be an approximation of the Burger dataset. In all cases, only users with more than $1,000$ tweets were included.

We report the key features of interest in Table 1. The *% gender-specific names* attribute is the percent of users whose first name had a greater than 90% association with a specific gender. The *% census-matched names* attribute is the percent of users whose first name was present in the census dataset. What is immediately apparent from these results is that our dataset maintains a nearly identical gender-associated and census-matched name composition as the random user set while the Burger dataset shows a significant deviation. This is evidence that where names are concerned (the focal point of the present work), our dataset construction method yields a sample of users accurately represents the general Twitter population. At the same time, the deviation of the Burger dataset from the random population confirms that this dataset (and therefore the results derived from it) are not representative of the general Twitter population.

We identify the general question of thoroughly quantifying sampling bias in gender (and other feature) labeled dataset construction as an important topic for additional work.

Gender inference methods

We evaluated the inference accuracy of three different, but related gender inference methods. All three used the SVM classifier and feature set that we have used in prior work (Liu, Zamal, and Ruths 2012; Zamal, Liu, and Ruths 2012). The key different between the three methods concerned the use of the user's name. In the first method, called *baseline*, name information was omitted entirely. In the second method, called *integrated*, the gender-association score of the name was added to the user's feature vector as a separate element (thus the integrated classifier used one more feature than the baseline classifier). The third method, called *threshold*, used the gender-association score of the user's name to decide whether to use the SVM-based classifier at all.

The core SVM classifier features

As the goal of this study was to determine the incremental value of using the user name as a feature in gender inference, we chose a standard set of features to give the SVM that have been used extensively in prior work (and have shown to perform quite well) (Liu, Zamal, and Ruths 2012; Zamal, Liu, and Ruths 2012; Pennacchiotti and Popescu 2011; Rustagi et al. 2009; Burger, Henderson, and Zarrella 2011).

k-top words. The k most differentiating words[3] used by each labeled group were included as individual features. To identify these k words in a corpus of tweets, every word, w, occurring in the corpus was assigned a score of $s(w) = u_1(w) - u_2(w)$, where $u_1(w)$ is the number of times the word appears in tweets generated by users with label 1, similarly for $u_2(w)$. The k words with the most negative and the k words with the most positive values were taken. Note that our formulation of k-top words produces $2k$ features — k top differentiating words for label 1 users and k top differentiating words for label 2 users. This principle applied to all k-top features below.

To compute the feature value for a k-top word, w, for a specific user, we evaluate to the frequency with which the word appeared in the user's microblog posts:

$$\frac{\text{\# of occurrences of } w}{\text{word count of the user's microblog stream}}.$$

This general formula (occurrence of the term of interest vs. occurrence of all terms of that type) was used for all k-top features that follow.

k-top stems. Plurals and verb forms can weaken the signal obtained from k-top words by causing forms of the same word to be handled as separate words (e.g., "houses", "housing", and "house" are all derived from the stem

[3]Tweets were converted to lowercase, with URLs and mentions and hashtags stripped out, then split on spaces.

"hous"). To address this, we passed all words through the Lovins stemmer and obtained the k-top differentiating stems for each labeled group, using the measure described above (Lovins 1968).

k-top digrams. In the training data, the k most differentiating digrams were identified for both labels exactly as described above, substituting digrams (two character sequences with spaces preserved) for words.

k-top trigrams. As with the words and digrams, the k-top trigrams (three character sequences), inclusive of spaces, were identified and included as features.

k-top co-stems. In prior work, the ends of words (e.g., conjugations, plurals, and possessive marks) were shown to give notable signal about a variety of blog author attributes. These strings, called *co-stems*, can be obtained by subtracting the stem returned by the Lovins stemmer and processing only the ending that remains (i.e., the word minus the stem). Co-stems were scored as described above and the k most differentiating co-stems for each label were added as features.

k-top hashtags. Hashtags operate as topic labels. Since certain topics may statistically align with gender labels, hashtags may be useful to the classifier. We applied the same scoring and selection approach as described above for words to identifying and quantifying differentiating hashtags for each label.

Frequency statistics. We also included a number of frequency statistics: tweets, mentions, hashtags, links, and retweets per day.

Retweeting tendency. The extent to which an individual propagates information was included by computing the ratio of retweets to tweets.

Neighborhood size. The ratio between number of followers and number of friends has been used as a measure of a user's tendency towards producing vs. consuming information on Twitter. We incorporated this as a feature as well.

The baseline classifier used precisely the set of features described above. We used an established SVM library, `libSVM`, as the classifier implementation (Chang and Lin 2011). After some experimentation, we chose the radial basis function as the SVM kernel with cost and gamma parameters chosen using a grid search technique.

The integrated classifier

When considering strategies for incorporating the user's name information into the classifier, the most obvious approach involved adding some name-specific elements to the user's feature vector. While prior work has chosen to encode names as a set of n-gram frequency features, we chose a different approach which dramatically decreased the number of name features added. Specifically, we added a single feature, the gender-name association score. The primary motivation behind this was to build in a priori knowledge (the strength of the association between a name and a gender label) that could significantly boost the performance of the classifier. A secondary motivation was to avoid overcomplicating the model representation of the user: using n-gram features (for $n > 1$) can quickly yield thousands of features - most of which will likely carry little information.

The threshold classifier

One potential weakness of the integrated classifier is that, as a feature, the gender-name association score will receive uniform weight in the classification problem, regardless of how strong the association is. This configuration is somewhat at odds with an intuitive interpretation of the association score: if a name is 100% associated with a specific gender, it should be incredibly hard for other features to override this signal. The threshold classifier implements the extreme version of this idea by setting a threshold for the gender-name association score; users with names having a gender association above this threshold are automatically given the appropriate gender label and the SVM classifier is skipped over entirely. If the association is below the threshold (i.e., the association is not sufficiently strong to trigger a quick decision), then the integrated classifier is run and the label returned by the SVM is applied to the user.

In using this classifier, the threshold value must be set prior to performing gender inferences. While seemingly an open-ended choice, there actually is a clear best value for the threshold: the average accuracy of the integrated classifier (when applied on its own). The intuition behind this choice involves recognizing that, by setting the threshold at τ, one is effectively accepting that for every name, x, with name-gender association, $\alpha_x \geq \tau$, the inferred label will be wrong $1-\alpha_x$ percent of the time. When is this an acceptable error rate? When the error rate of the association-based inference is better than the error rate of the alternative—in this case the integrated classifier.

Thus, since we can characterize the error rate, ϵ, of the integrated classifier, a natural choice for the threshold is $\tau = 1 - \epsilon$. Any association-based inference using a name with $\alpha_x > \tau$ will be right more often than the integrated classifier.

Note that, in theory, we can do better than even this if we have some additional information. What we actually want to ensure is that the effective error rate of *all* association-based inferences is less than the error rate of the integrated classifier. Since every name, x, will occur in the population with a specific probability, p_x, the actual error rate of the association-based component of the classifier is $1 - A_\tau$, where

$$A_\tau = \sum_{x \in X_\tau} p_x \alpha_x$$

and $X_\tau = \{x_1, x_2, ...\}$ is the set of all names with a gender-name association $\alpha_x > \tau$. The best choice of τ, then is the value for which $A_\tau = \epsilon$. Observe that the earlier choice of $\tau = \epsilon$ is a conservative estimate since this assumes that $p_x > 0$ only for names that have $\alpha_x = \epsilon$. In general,

Table 2: The accuracy of the three methods when run on the profile picture-based Twitter dataset. Both name-based methods show notable improvement over the baseline. The threshold-based system shows the most significant increase in accuracy.

Method	*Male Acc*	*Female Acc*	*Avg Acc*
Baseline	83.6	83.0	**83.3**
Integrated	86.0	84.3	**85.2**
Threshold, $\tau = 1.0$	86.4	86.3	86.4
Threshold, $\tau = 0.90$	87.4	86.6	87.0
Threshold, $\tau = 0.85$	87.5	86.6	**87.1**
Threshold, $\tau = 0.70$	87.5	86.6	87.1

the more highly gender-associated names that occur in a population, the lower the choice of τ can be.

Results and Discussion

In order to evaluate the relative performance of the three methods (baseline, integrated, and threshold), each was run on the dataset constructed from user profile pictures. 10-fold cross validation (400 males and 400 females per fold) was used to measure the performance of each method. Table 2 shows the results of this exercise for all three methods. Several trends are noteworthy.

The addition of name information improves accuracy Both methods that incorporate name information outperform the baseline. Crucially, besides the work of (Burger, Henderson, and Zarrella 2011) (which we have explained as a special case), both methods outperform all gender inference systems with which we are familiar.

Figure 1 shows the distribution of names by gender association. The tri-modal characteristics of the distribution reveals why the gender association of a name is helpful to the inference problem: most names are either strongly associated with a given gender or unknown. Furthermore, the extreme bias towards strongly gender-associated words also explains why the threshold approach makes only a modest improvement on the integrated method. The idea behind the threshold approach was to correct for situations in which the integrated method would not weight the gender-association feature strongly enough in making a classification. However, Figure 1 reveals that most names have either strong gender-association or no association. As a result, this mis-weighting situation will only occur when handling the relatively few users who have names that fall between these peaks. This said, since the gender-associations are so strong, using the threshold-based approach may generally outperform the integrated method since in the threshold-based method, there will be no other features to confound the gender label implied by a strongly gender-associated name.

Unknown names come in different forms. The tall central peak around zero in Figure 1 indicates that well more than half of the users (66%) have a name that has an unknown gender association. Manual inspection of

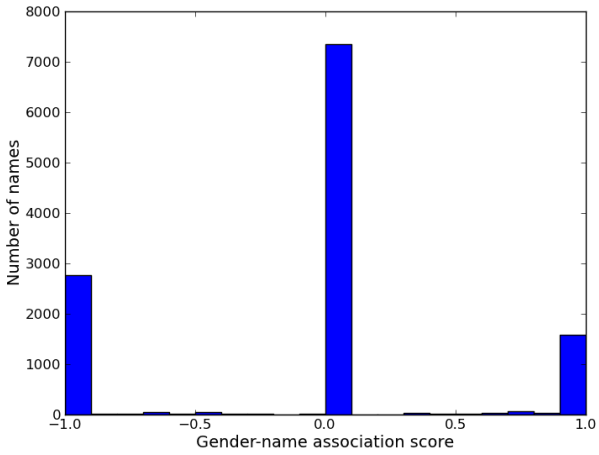

Figure 1: The distribution of names in the dataset by gender-name association. Of particular note is the tri-modal characteristic of the distribution, suggesting that most names are either strongly male, female, or unknown.

strings that are not handled reveals two distinct types. One type consists of well-formed names whose first name is simply not contained in the dictionary of known names (e.g., "Lim" and "Faizan"). However, the much larger class are ill-formed name strings. These are of at least four subtypes:

- nicknames and name abbreviations (e.g., "Big Daddy C", "BIG_SHAWN_20" and "CJ Sullivan"): we consider these strings ill-formed names on the grounds that their vocabulary does not intersect with strings that are considered full name elements.
- mangled names (e.g., "AlanLeong" and "[-!Raphael-]"): these are names which would be valid except that they are run together, decorated with non-standard characters, or made otherwise difficult to parse due to formatting decisions.
- usernames (strings that read like a Twitter username or email address, e.g., "swagboiboo504"), and
- non-names (e.g., "Married To Bieber ;)", "The Great Gatsby", and "25 MORE DAYS"): these are strings which are not intended to function as a name at all. Typical strings belonging in this category are primarily status updates and music/celebrity references.

This categorization of ill-formed name strings suggests that there remains significant opportunities for mining gender signal from the name field. As illustrated in the examples given above, nicknames, abbreviations, mangled names, and usernames can frequently contain non-trivial gender cues. Identifying strategies for extracting and using these cues to more accurately infer gender is a promising direction for future work.

In terms of immediate next steps with this research direction, it is helpful to recognized that the gender-name

association has an inherent weakness - specifically that names which were not seen in the training data give no a priori knowledge to the classifier. Certainly this is an issue, particularly if the classifier were to be extended to more diverse (e.g., different language/region) populations in which name diversity might increase. Despite this, we still consider the gender-name association a promising approach since the measure compactly encapsulates a prior on the gender of the individual based on their name. This is something that is obtained nearly for free which, we suspect, would require a tremendous amount of training data to reproduce using n-gram-based features. With this in mind, we consider two promising directions for extending the present effort.

First, one might consider ways of obtaining more expansive a priori knowledge (e.g., gender-name association data) from 3rd party sources. Any platform or dataset in which gender is explicitly indicated would provide a valuable source of information for augmenting the associations we derived from US census data.

In parallel to this investigation, it would be productive to consider combining the n-gram feature based model advanced by Burger et al. with our gender-name association approach. Since both provide signal of the gender label, presumably the combination would perform as well or better.

Conclusions

In this paper, we have presented two novel methods for leveraging a Twitter user's self-reported name in order to infer her gender. The methods outperform all gender inference methods currently available or proposed for Twitter. Beyond the absolute performance of these methods, we hope that our work calls attention to the name field (as well as other metadata) that may help the latent attribute inference problem. Despite its obvious information content, the name field has been, to date, remarkably underutilized. Our work demonstrates the performance gain that can be obtained through its inclusion in the classification task.

In addition to these inference methods, we have also devised a new strategy for generating gender-labeled Twitter datasets without depending on textual content from the user in any way. The resulting datasets reflect aspects of the general Twitter population well—an important fact if we are to trust accuracy estimates on the datasets to generalize to Twitter in general.

Finally, we have released the gender-labeled Twitter dataset used as part of this study for the community. Our hope is that this is the first of many datasets that will be made available for the community, enabling more comparative analysis of latent attribute inference methods on Twitter.

References

Buhrmester, M.; Kwang, T.; and Gosling, S. D. 2011. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science* 6(1):3–5.

Burger, J.; Henderson, J.; and Zarrella, G. 2011. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27.

Conover, M.; Gonalves, B.; Ratkiewicz, J.; Flammini, A.; and Menczer, F. 2011. Predicting the politial alignment of twitter users. In *Proceedings of the International Conference on Social Computing*.

Littlestone, N. 1988. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning* 2:285–318.

Liu, W.; Zamal, F. A.; and Ruths, D. 2012. Using social media to infer gender composition from commuter populations. In *Proceedings of the When the City Meets the Citizen Workshop, the International Conference on Weblogs and Social Media*.

Lovins, J. 1968. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics* 11(1):21–31.

Mislove, A.; Lehmann, S.; Ahn, Y.; Onnela, J.; and Rosenquist, J. 2011. Understanding the Demographics of Twitter Users. In *Proceedings of the International Conference on Weblogs and Social Media*.

Pennacchiotti, M., and Popescu, A. 2011. A machine learning approach to twitter user classification. In *Proceedings of the International Conference on Weblogs and Social Media*.

Rustagi, M.; Prasath, R.; Goswami, S.; and Sarkar, S. 2009. Learning age and gender of blogger from stylistic variation. In *Proceedings of the International Conference on Pattern Recognition and Machine Learning*.

Schnoebelen, T., and Kuperman, V. 2010. Using amazon mechanical turk for linguistic research. *Psihologija* 43(4):441–464.

Zamal, F. A.; Liu, W.; and Ruths, D. 2012. Homophily and latent attribute inference: inferring latent attributes of Twitter users from neighbors. In *Proceedings of the International Conference on Weblogs and Social Media*.