WHAT'S IN A NAME?

Using first names as features for gender inference in Twitter

Wendy Liu
wendy.liu@mail.mcgill.ca

Derek Ruths
derek.ruths@mcgill.ca

Network Dynamics Lab (www.networkdynamics.org)
School of Computer Science / McGill University / Montreal, Canada

Introduction

Despite significant work on the problem of inferring a Twitter user's gender from her online content, no systematic research has been done on using the most obvious signal of gender: first name. In this study, we investigate the link between gender and first name in English tweets.

Methods

Building the gender-labelled dataset

As there are no canonical gender-labelled Twitter datasets available to the research community, we developed our own for this study. The process of obtaining the labels is illustrated in Figure 1. Note that this process does not make use of the textual content of a user in any way. We then collected the 1,000 most recent tweets for each labelled user.

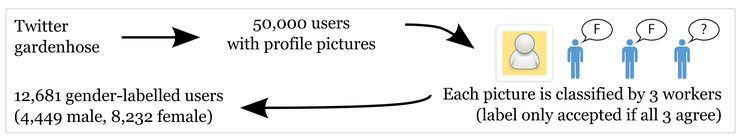

Figure 1. Using Amazon Mechanical Turk to build a gender-labelled dataset.

To ensure that our method resulted in a representative sample of users, we computed several statistics over our dataset, as well as over 100,000 randomly selected English users and a dataset constructed similarly to that used in Burger et al, 2011 (since the original was not available). Figure 2 shows that the Burger dataset is unlikely to be representative of the general Twitter population, whereas ours yields a more representative sample, especially where names are concerned. We note that our method also eliminates any correlation between labels and user content that might artificially inflate reported accuracy.

Dataset	Average tweets / user	% gender-associated names (>90% association)	% census-matched names
Random	13,221	34.1%	37.3%
LiuRuths	15,948	33.8%	37.1%
Burger	18,385	27.5%	29.4%

Figure 2. Attributes of our dataset vs. random tweets and an approximation of Burger et al.

Gender inference methods

We evaluated 3 gender inference systems on this gold standard dataset, subsampled to obtain 4,000 users per gender:

The baseline classifier. With this method, name information was omitted entirely. Classification was carried out using the `libSVM` implementation of a support vector machine, using the features shown in Figure 3, after which 10-fold cross validation was performed.

k-top differentiating features for each gender (k=20): hashtags ('#hashtag'), words, digrams, trigrams, stems ('emerg'), costems ('ency')
Frequency features (per day): tweets, retweets, hashtags, mentions
Ratio features: tweet:retweet, followers:followees

Figure 3. Features studied per user.

The integrated classifier. The gender association score of the user's first name, g(x), was computed using name distribution data in the 1990 US census, then used as an additional feature for the SVM classifier:

$$g(x) = \frac{M(x) - F(x)}{M(x) + F(x)}$$

where M(x) and F(x) are the number of males and females with that first name, respectively. The score ranges from -1 (only given to females) to 1 (only given to males). A name not in the dictionary is given a score of 0, indicating no a priori knowledge.

The threshold classifier. Rather than using g(x) as a feature for the feature vector, a threshold of τ was set such that whenever g(x) > τ, the gender label associated with g(x) is accepted as the label and the SVM-based classification is skipped entirely. Otherwise, the integrated classifier is used to determine a label.

Results and discussion

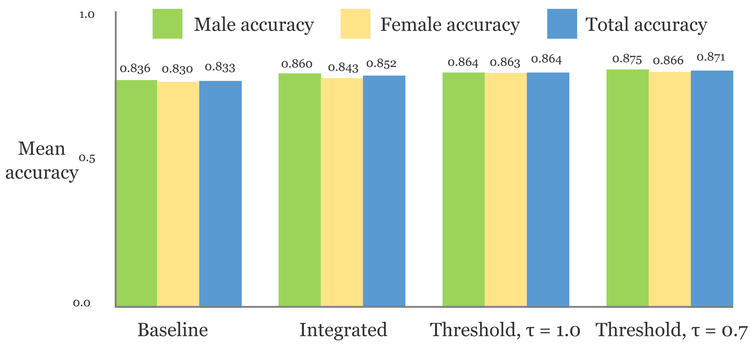

Figure 4. SVM classifier results for all methods.

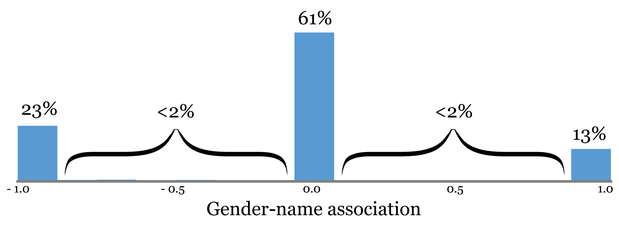

Figure 5. Distribution of first names of Twitter users by US census gender score.

The addition of name information improves accuracy. Both methods that incorporate name information outperform the baseline (Figure 4). Indeed, besides the work of Burger et al, which is shown in Figure 1 to be a special case, both methods outperform all gender inference systems with which we are familiar. The tri-modal characteristics of the name distribution (Figure 5) reveal why: the vast majority of first names are either strongly associated with a specific gender, or unknown. Thus names of the first form will give off a strong signal, which will agree with AMT-provided labels 99% of the time.

However, using name information alone is not likely to result in 99% accuracy on an arbitrary dataset, as the tall central peak in Figure 5 shows; the majority of users in this dataset have a first name with an unknown gender association. This indicates that name information can only be successfully used in conjunction with another inference method, such as the baseline method used in this study.

Unknown names come in different forms. Manual inspection of first names located in the central peak of Figure 5 reveals several distinct types, illustrated with examples in Figure 6. Such names can still contain non-trivial gender cues. Identifying strategies for extracting and using these cues is a promising direction for future work.

Names not in census data	Nicknames and abbreviations	Mangled names	Non-names
Lim	CJ Sullivan	AlanLeong	Married To Bieber :)
Faizan	Big Daddy C	[-!Raphael!-]	25 MORE DAYS
Saleem	J.T.	PeterAzP	NoExcuses

Figure 6. The various types of names with a score of 0, with examples.

Conclusion

In this paper, we presented two novel methods for incorporating the user's self-reported name into a gender classifier. This yields a 20% increase in accuracy over a standard baseline classifier.

In addition to these inference methods, we developed a novel way of obtaining gender labels that does not require analysis of the user's textual information, and have built a large dataset of gender-labelled Twitter users which we have published for community use (download it from http://bit.ly/microtext2013). Our hope is that this will provide researchers with a basis for comparison of gender inference methods.

References

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