Discriminating Gender on Twitter

John D. Burger and **John Henderson** and **George Kim** and **Guido Zarrella**
The MITRE Corporation
202 Burlington Road
Bedford, Massachusetts, USA 01730
{john,jhndrsn,gkim,jzarrella}@mitre.org

Abstract

Accurate prediction of demographic attributes from social media and other informal online content is valuable for marketing, personalization, and legal investigation. This paper describes the construction of a large, multilingual dataset labeled with gender, and investigates statistical models for determining the gender of uncharacterized Twitter users. We explore several different classifier types on this dataset. We show the degree to which classifier accuracy varies based on tweet volumes as well as when various kinds of profile metadata are included in the models. We also perform a large-scale human assessment using Amazon Mechanical Turk. Our methods significantly out-perform both baseline models and almost all humans on the same task.

1 Introduction

The rapid growth of social media in recent years, exemplified by Facebook and Twitter, has led to a massive volume of user-generated informal text. This in turn has sparked a great deal of research interest in aspects of social media, including automatically identifying latent demographic features of online users. Many latent features have been explored, but gender and age have generated great interest (Schler et al., 2006; Burger and Henderson, 2006; Argamon et al., 2007; Mukherjee and Liu, 2010; Rao et al., 2010). Accurate prediction of these features would be useful for marketing and personalization concerns, as well as for legal investigation.

In this work, we investigate the development of high-performance classifiers for identifying the gender of Twitter users. We cast gender identification as the obvious binary classification problem, and explore the use of a number of text-based features. In Section 2, we describe our Twitter corpus, and our methods for labeling a large subset of this data for gender. In Section 3 we discuss the features that are used in our classifiers. We describe our Experiments in Section 4, including our exploration of several different classifier types. In Section 5 we present and analyze performance results, and discuss some directions for acquiring additional data by simple self-training techniques. Finally in Section 6 we summarize our findings, and describe extensions to the work that we are currently exploring.

2 Data

Twitter is a social networking and micro-blogging platform whose users publish short messages or *tweets*. In late 2010, it was estimated that Twitter had 175 million registered users worldwide, producing 65 million tweets per day (Miller, 2010). Twitter is an attractive venue for research into social media because of its large volume, diverse and multilingual population, and the generous nature of its Terms of Service. This has led many researchers to build corpora of Twitter data (Petrovic et al., 2010; Eisenstein et al., 2010). In April 2009, we began sampling data from Twitter using their API at a rate of approximately 400,000 tweets per day. This represented approximately 2% of Twitter's daily volume at the time, but this fraction has steadily decreased to less than 1% by 2011. This decrease is because we sample roughly the same number of tweets every day while Twitter's overall volume has increased markedly. Our corpus thus far contains approximately 213 million tweets from 18.5 million users, in many different languages.

In addition to the tweets that they produce, each Twitter user has a profile with the following free-text fields:

- Screen name (e.g., *jsmith92*, *kingofpittsburgh*)
- Full name (e.g., *John Smith*, *King of Pittsburgh*)
- Location (e.g., *Earth*, *Paris*)
- URL (e.g., the user's web site, Facebook page, etc.)
- Description (e.g., *Retired accountant and grandfather*)

All of these except screen name are completely optional, and all may be changed at any time. Note that none

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 1301–1309,
Edinburgh, Scotland, UK, July 27–31, 2011.

	Users	Tweets
Training	146,925	3,280,532
Development	18,380	403,830
Test	18,424	418,072

Figure 1: Dataset Sizes

of the demographic attributes we might be interested in are present, such as gender or age. Thus, the existing profile elements are not directly useful when we wish to apply supervised learning approaches to classify tweets for these target attributes. Other researchers have solved this problem by using labor-intensive methods. For example, Rao et al. (2010) use a focused search methodology followed by manual annotation to produce a dataset of 500 English users labeled with gender. It is infeasible to build a large multilingual dataset in this way, however.

Previous research into gender variation in online discourse (Herring et al., 2004; Huffaker, 2004) has found it convenient to examine blogs, in part because blog sites often have rich profile pages, with explicit entries for gender and other attributes of interest. Many Twitter users use the URL field in their profile to link to another facet of their online presence. A significant number of users link to blogging websites, and many of these have well-structured profile pages indicating our target attributes. In many cases, these are not free text fields. Users on these sites must select gender and other attributes from drop-down menus in order to populate their profile information. Accordingly, we automatically followed the Twitter URL links to several of the most represented blog sites in our dataset, and sampled the corresponding profiles. By attributing this blogger profile information to the associated Twitter account, we created a corpus of approximately 184,000 Twitter users labeled with gender.

We partitioned our dataset by user into three distinct subsets, training, development, and test, with sizes as indicated in Figure 1. That is, all the tweets from each user are in a single one of the three subsets. This is the corpus we use in the remainder of this paper.

This method of gleaning supervised labels for our Twitter data is only useful if the blog profiles are in turn accurate. We conducted a small-scale quality assurance study of these labels. We randomly selected 1000 Twitter users from our training set and manually examined the description field for obvious indicators of gender, e.g., *mother to 3 boys* or *just a dude*. Only 150 descriptions (15% of the sample) had such an explicit gender cue. 136 of these also had a blog profile with the gender selected, and in all of these the gender cue from the user's Twitter description agreed with the corresponding blog profile. This may only indicate that people who misrepresent their gender are simply consistent across different aspects of their online presence. However, the effort involved in maintaining this deception in two different places suggests that the blog labels on the Twitter data are largely reliable.

Initial analysis using the blog-derived labels showed that our corpus is composed of 55% females and 45% males. This is consistent with the results of an earlier study which used name/gender correlations to estimate that Twitter is 55% female (Heil and Piskorski, 2009). Figure 2 shows several statistics broken down by gender, including the Twitter users who did not indicate their gender on their blog profile. In our dataset females tweet at a higher rate than males and in general users who provide their gender on their blog profile produce more tweets than users who do not. Additionally, of the 150 users who provided a gender cue in their Twitter user description, 105 were female (70%). Thus, females appear more likely to provide explicit indicators about their gender in our corpus.

The average number of tweets per user is 22 and is fairly consistent across our traing/dev/test splits. There is wide variance, however, with some users represented by only a single tweet, while the most prolific user in our sample has nearly 4000 tweets.

It is worth noting that many Twitter users do not tweet in English. Table 3 presents an estimated breakdown of language use in our dataset. We ran automatic language ID on the concatenated tweet texts of each user in the training set. The strong preponderance of English in our dataset departs somewhat from recent studies of Twitter language use (Wauters, 2010). This is likely due in part to sampling methodology differences between the two studies. The subset of Twitter users who also use a blog site may be different from the Twitter population as a whole, and may also be different from the users tweeting during the three days of Wauters's study. There are also possible longitudinal differences: English was the dominant language on Twitter when the online service began in 2006, and this was still the case when we began sampling tweets in 2009, but the proportion of English tweets had steadily dropped to about 50% in late 2010. Note that we do not use any explicit encoding of language information in any of the experiments described below.

Our Twitter-blog dataset may not be entirely representative of the Twitter population at general, but this has at least one advantage. As with any part of the Internet, spam is endemic to Twitter. However by sampling only Twitter users with blogs we have largely filtered out spammers from our dataset. Informal inspection of a few thousand tweets revealed a negligible number of commercial tweets.

3 Features

Tweets are tagged with many sources of potentially discriminative metadata, including timestamps, user color

	Users		Tweets		Mean tweets
	Count	Percentage	Count	Percentage	per user
Female	100,654	42.3%	2,429,621	47.7%	24.1
Male	83,075	35.0	1,672,813	32.8	20.1
Not provided	53,817	22.7	993,671	19.5	18.5

Figure 2: Gender distribution in our blog-Twitter dataset

Language	Users	Percentage
English	98,004	66.7%
Portuguese	21,103	14.4
Spanish	8,784	6.0
Indonesian	6,490	4.4
Malay	1,401	1.0
German	1,220	0.8
Chinese	985	0.7
Japanese	962	0.7
French	878	0.6
Dutch	761	0.5
Swedish	686	0.5
Filipino	643	0.4
Italian	631	0.4
Other	4,377	3.0

Figure 3: Language ID statistics from training set

preferences, icons, and images. We have restricted our experiments to a subset of the textual sources of features as listed in Figure 4.

We use the content of the tweet text as well as three fields from the Twitter user profile described in Section 2: full name, screen name, and description. For each user in our dataset, a field is in general a *set* of text strings. This is obviously true for tweet texts but is also the case for the profile-based fields since a Twitter user may change any part of their profile at any time. Because our sample spans points in time where users have changed their screen name, full name or description, we include all of the different values for those fields as a set. In addition, a user may leave their description and full name blank, which corresponds to the empty set.

In general, our features are quite simple. Both word- and character-level ngrams from each of the four fields are included, with and without case-folding. Our feature functions do not count multiple occurrences of the same ngram. Initial experiments with count-valued feature functions showed no appreciable difference in performance. Each feature is a simple Boolean indicator representing presence or absence of the word or character ngram in the set of text strings associated with the particular field. The extracted set of such features represents the item to the classifier.

For word ngrams, we perform a simple tokenization

	Feature extraction		
	Char ngrams	Word ngrams	Distinct features
Screen name	1–5	*none*	432,606
Full name	1–5	1	432,820
Description	1–5	1–2	1,299,556
Tweets	1–5	1–2	13,407,571
Total			15,572,522

Figure 4: Feature types and counts

that separates words at transitions between alphanumeric characters and non-alphanumeric.[1] We make no attempt to tokenize unsegmented languages such as Chinese, nor do we perform morphological analysis on language such as Korean; we do no language-specific processing at all. We expect the character-level ngrams to extract useful information in the case of such languages.

Figure 4 indicates the details and feature counts for the fields from our training data. We ignore all features exhibited by fewer than three users.

4 Experiments

We formulate gender labeling as the obvious binary classification problem. The sheer volume of data presents a challenge for many of the available machine learning toolkits, e.g. WEKA (Hall et al., 2009) or MALLET (McCallum, 2002). Our 4.1 million tweet training corpus contains 15.6 million distinct features, with feature vectors for some experiments requiring over 20 gigabytes of storage. To speed experimentation and reduce the memory footprint, we perform a one-time feature generation preprocessing step in which we convert each feature pattern (such as "caseful screen name character trigram: Joh") to an integer codeword. The learning algorithms do not access the codebook at any time and instead deal solely with vectors of integers. We compress the data further by concatenating all of a user's features into a single vector that represents the union of every tweet produced by that user. This condenses the dataset to about 180,000 vectors occupying 11 gigabytes of storage.

We performed initial feasibility experiments using a wide variety of different classifier types, including Support Vector Machines, Naive Bayes, and Balanced Win-

[1] We use the standard regular expression pattern `\b`.

now2 (Littlestone, 1988). These initial experiments were based only on caseful word unigram features from tweet texts, which represent less than 3% of the total feature space but still include large numbers of irrelevant features. Performance as measured on the development set ranged from Naive Bayes at 67.0% accuracy to Balanced Winnow2 at 74.0% accuracy. A LIBSVM (Chang and Lin, 2001) implementation of SVM with a linear kernel achieved 71.8% accuracy, but required over fifteen hours of training time while Winnow needed less than seven minutes. No classifier that we evaluated was able to match Winnow's combination of accuracy, speed, and robustness to increasing amounts of irrelevant features.

We built our own implementation of the Balanced Winnow2 algorithm which allowed us to iterate repeatedly over the training data on disk rather than caching the entire dataset in memory. This reduced our memory requirements to the point that we were able to train on the entire dataset using a single machine with 8 gigabytes of RAM.

We performed a grid search to select learning parameters by measuring their affect on Winnow's performance on the development set. We found that two sets of parameters were required: a low learning rate (0.03) was effective when using only one type of input feature (such as only screen name features, or only tweet text features), and a higher learning rate (0.20) was required when mixing multiple types of features in one classifier. In both cases we used a relatively large margin (35%) and cooled the learning rate by 50% after each iteration.

These learning parameters were used during all of the experiments that follow. All gender prediction models were trained using data from the training set and evaluated on data from the development set. The test set was held out entirely until we finalized our best performing models.

4.1 Field combinations

We performed a number of experiments with the Winnow algorithm described above. We trained it on the training set and evaluated on the development set for each of the four user fields in isolation, as well as various combinations, in order to simulate different use cases for systems that perform gender prediction from social media sources. In some cases we may have all of the metadata fields available above, while in other cases we may only have a sample of a user's tweet content or perhaps just one tweet. We simulated the latter condition by randomly selecting a single tweet for each dev and test user; this tweet was used for all evaluations of that user under the single-tweet condition. Note, however, that for training the single tweet classifier, we do not concatenate all of a user's tweets as described above. Instead, we pair each user in the training set with each of their tweets in turn, in order to take advantage of all the training data. This amounted to over 3 million training instances for the single tweet condition.

We paid special attention to three conditions: single tweet, all fields, and all tweets. For these conditions, we evaluated the learned models on the training data, the development set, and the test set, to study over-training and generalization. Note that for all experiments, the evaluation includes some users who have left their full name or description fields blank in their profile.

In all cases, we compare results to a maximum likelihood baseline that simply labels all users female.

4.2 Human performance

We wished to compare our classifier's efficacy to human performance on the same task. A number of researchers have recently experimented with the use of Amazon Mechanical Turk (AMT) to create and evaluate human language data (Callison-Burch and Dredze, 2010). AMT and other crowd-sourcing platforms allow simple tasks to be posted online for large numbers of anonymous workers to complete.

We used AMT to measure human performance on gender determination for the all tweets condition. Each AMT worker was presented with all of the tweet texts from a single Twitter user in our development set and asked whether the author was male or female. We redundantly assigned five workers to each Twitter user, for a total of 91,900 responses from 794 different workers. We experimented with a number of ways to combine the five human labels for each item, including a simple majority vote and a more sophisticated scheme using an expectation maximization algorithm.

4.3 Self-training

Our final experiments were focused on exploring the use of unlabeled data, of which we have a great deal. We performed some initial experiments on a self-training approach to labeling more data. We trained the all-fields classifier on half of our training data, and applied it to the other half. We trained a new classifier on this full training set, which now included label errors introduced by the limitations of the first classifier. This provided a simulation of a self-training setup using half the training data. Any robust gains due to self-training should be revealed by this setup.

5 Results

5.1 Field combinations

Figure 5 shows development set performance on various combinations of the user fields, all of which outperform the maximum likelihood baseline that classifies all users as female. The single most informative field with respect

Baseline (F)	54.9%
One tweet text	67.8
Description	71.2
All tweet texts	75.5
Screen name (e.g. *jsmith92*)	77.1
Full name (e.g. *John Smith*)	89.1
Tweet texts + screen name	81.4
Tweet texts + screen name + description	84.3
All four fields	92.0

Figure 5: Development set accuracy using various fields

Condition	Train	Dev	Test
Baseline (F)	54.8%	54.9	54.3
One tweet text	77.8	67.8	66.5
Tweet texts	77.9	75.5	74.5
All fields	98.6	92.0	91.8

Figure 6: Accuracy on the training, development and test sets

to gender is the user's full name, which provides an accuracy of 89.1%. Screen name is often a derivative of full name, and it too is informative (77.1%), as is the user's self-assigned description (71.2).

Using only tweet texts performs better than using only the user description (75.5% vs. 71.2). Tweet texts are sufficient to decrease the error by nearly half over the all-female prior. It appears that the tweet texts convey more about a Twitter user's gender than their own self-descriptions. Even a single (randomly selected) tweet text contains some gender-indicative information (67.2%). These results are similar to previous work. Rao et al. (2010) report results of 68.7% accuracy on gender from tweet texts alone using an ngram-only model, rising to 72.3 with hand-crafted "sociolinguistic-based" features. Test set differences aside, this is comparable with the "All tweet texts" line in Figure 5, where we achieve an accuracy of 75.5%.

Performance of models built from various aggregates of the four basic fields are shown in Figure 5 as well. The combination of tweet texts and a screen name represents a use case common to many different social media sites, such as chat rooms and news article comment streams. The performance of this combination (81.4%) is significantly higher than either of the individual components. As we have observed, full name is the single most informative field. It out-performs the combination of the other three fields, which perform at 84.3%. Finally, the classifier that has access to features from all four fields is able to achieve an accuracy of 92.0%.

The final test set accuracy is shown in Figure 6. This test set was held out entirely during development and has been evaluated only with the four final models reported

Rank	MI	Feature f	$P(Female\|f)$
1	0.0170	!	0.601
2	0.0164	_:	0.656
3	0.0163	_lov	0.687
4	0.0162	love	0.680
5	0.0161	lov	0.676
6	0.0160	_love	0.689
7	0.0160	!_	0.618
8	0.0149	:)	0.697
9	0.0148	y!	0.687
10	0.0145	**my**	0.637
11	0.0143	love_	0.691
12	0.0143	haha	0.705
13	0.0141	my_	0.634
14	0.0140	_my	0.637
15	0.0140	_:)	0.697
16	0.0139	_my	0.634
17	0.0138	!_i	0.711
18	0.0138	hah	0.698
19	0.0137	_hah	0.714
20	0.0135	_so	0.661
21	0.0134	_haha	0.714
22	0.0132	**so**	0.661
23	0.0128	_i	0.618
24	0.0127	ooo	0.708
25	0.0126	!_i	0.743
26	0.0123	i_lov	0.728
27	0.0120	ove_	0.671
28	0.0117	ay!	0.718
29	0.0116	aha	0.678
30	0.0116	<3	0.856
31	0.0115	_cute	0.826
32	0.0114	i_lo	0.704
33	0.0114	:)$	0.701
34	0.0110	:(	0.731
35	0.0109	_:)$	0.701
36	0.0109	!$	0.614
37	0.0107	ahah	0.716
38	0.0106	_<3	0.857
464	0.0051	_ht	♂ 0.506
465	0.0051	hank	0.641
466	0.0051	too_	0.659
467	0.0051	_yay!	0.818
468	0.0051	_http	♂ 0.506
469	0.0051	_htt	♂ 0.506
624	0.0047	Googl	♂ 0.317
625	0.0047	ing!_	0.718
626	0.0047	hair_	0.749
627	0.0047	_b	0.573
628	0.0047	y_:	0.725
629	0.0046	Goog	♂ 0.318

Figure 7: A selection of tweet text features, ranked by mutual information. Character ngrams in Courier, words in **bold**. Underscores are spaces, $ matches the end of the tweet text. ♂ marks "male" features.

in this figure. The difference between the scores on the train and development sets show how well the model can fit the data. There are features in the user name and user screen name fields that make the data trivially separable. The tweet texts, however, present more ambiguity for the learners. The difference between the development and test set scores suggest that only minimal hill-climbing occurred during our development.

We have performed experiments to better understand how performance scales with training data size. Figure 8 shows how performance increases for both the all-fields and tweet-texts-only classifiers as we train on more users, with little indication of leveling off.

As discussed in Section 2, there is wide variance in the number of tweets available from different users. In Figure 9 we show how the tweet text classifier's accuracy increases as the number of tweets from the user increases. Each point is the average classifier accuracy for the user cohort with exactly that many tweets in our dev set. Performance increases given more tweets, although the averages get noisy for the larger tweet sets, due to successively smaller cohort sizes.

Some of the most informative features from tweet texts are shown in Figure 7, ordered by mutual information with gender. There are far more of these strong features for the female category than the male: only five of the top 1000 features are associated more strongly with males, i.e. they have lower $P(Female|feature)$ than the prior, $P(Female) = 0.55$.

Some of these features are content-based (*hair*, and several fragments of *love*), while others are stylistic (*ooo*, several emoticons). The presence of *http* as a strong male feature might be taken to indicate that men include links in their tweet texts far more often than women, but a cursory examination seems to show instead that women are simply more likely to include "bare" links, e.g., `emnlp.org` vs. `http://emnlp.org`.

5.2 Human performance

Figure 10 shows the results of the human performance benchmarks using Amazon Mechanical Turk. The raw per-response performance is 60.4%, only moderately better than the all-female baseline. When averaged across workers, however, this improves substantially, to 68.7. This would seem to indicate that there were a few poor workers who did many annotations, and in fact when we limit the performance average to those workers who produced 100 or more responses, we do see a degradation to 62.2.

The problem of poor quality workers is endemic to anonymous crowd sourcing platforms like Mechanical Turk. A common way to combat this is to use redundancy, with a simple majority vote to choose among multiple responses for each item. This allows us to treat the five workers who responded to each item as an ensemble. As Figure 10 indicates, this provides some improvement over the raw result (65.7% vs. 60.4). A different approach, first proposed by Dawid and Skene (1979), is to use an expectation maximization algorithm to estimate the quality of each source of labels, as well as estimate the posterior for each item. In this case, the first is an AMT worker's capability and the second is the distribution of gender labels for each Twitter user.

Baseline	54.9
Average response	60.4
Average worker	68.7
Average worker (100 or more responses)	62.2
Worker ensemble, majority vote	65.7
Worker ensemble, EM-adjusted vote	67.3
Winnow all-tweet-texts classifier	75.5

Figure 10: Comparing with humans on the all tweet texts task

The Dawid and Skene approach has previously been applied to Mechanical Turk responses (Ipeirotis et al., 2010). We used their implementation on our AMT results but with only moderate improvement over the simple majority ensemble (67.3% vs. 65.7). All of the aggregate human results are substantially below the all-tweet-texts classifier score, suggesting that this is a difficult task for people to perform. As Figure 11 indicates, most workers perform below 80% accuracy, and less than 5% of the prolific workers out-perform the automatic classifier. These high-scoring workers may indeed be good at the task, or they may have simply been assigned a less-difficult subset of the data. Figure 12 illustrates this by showing aligned worker performance and classifier performance on the precise set of items that each worker performed on. Here we see that, with few exceptions, the automatic classifier performs as well or better than the AMT workers on their subset.

5.3 Self-training

Finally, as described in Section 4.3, we performed some initial experiments on a self-training approach to labeling more data. As described above the all-fields classifier achieves an accuracy of 92% on the development set when trained on the full training set. Training on half of the training data results in a drop to 91.1%. The second classifier trained on the full training set, but with some label errors introduced by the first, had further degraded performance of 90.9%. Apparently the errorful labels introduced by the simplistic self-training procedure overwhelmed any new information that might have been gained from the additional data. We are continuing to explore ways to use the large amounts of unsupervised data in our corpus.

Figure 8: Performance increases when training with more users

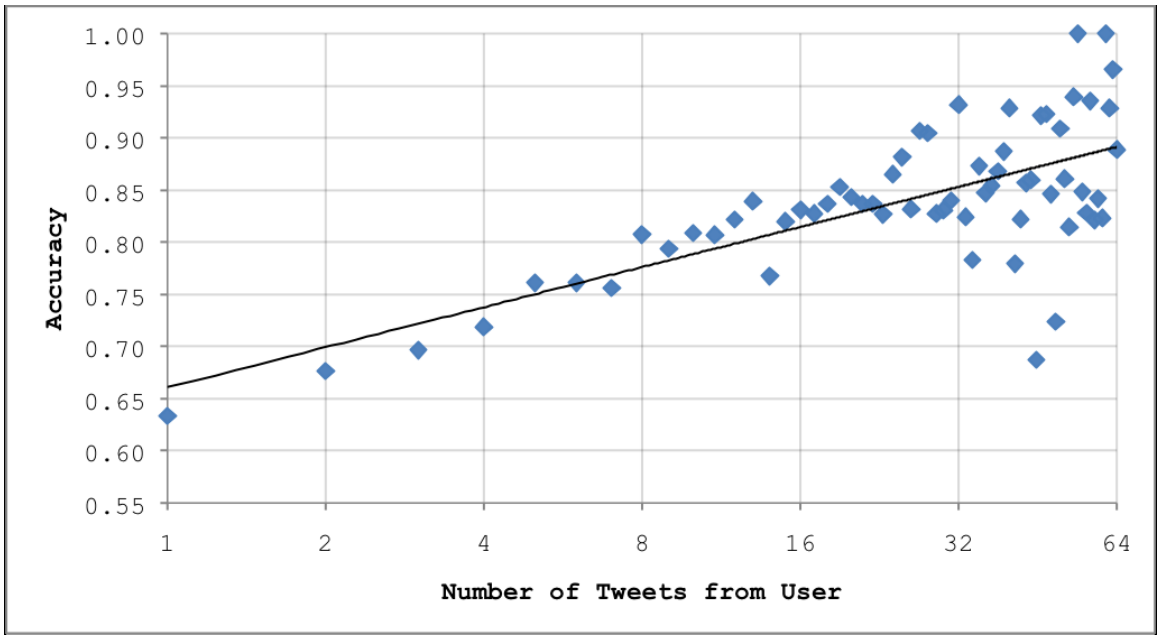

Figure 9: Performance increases with more tweets from target user

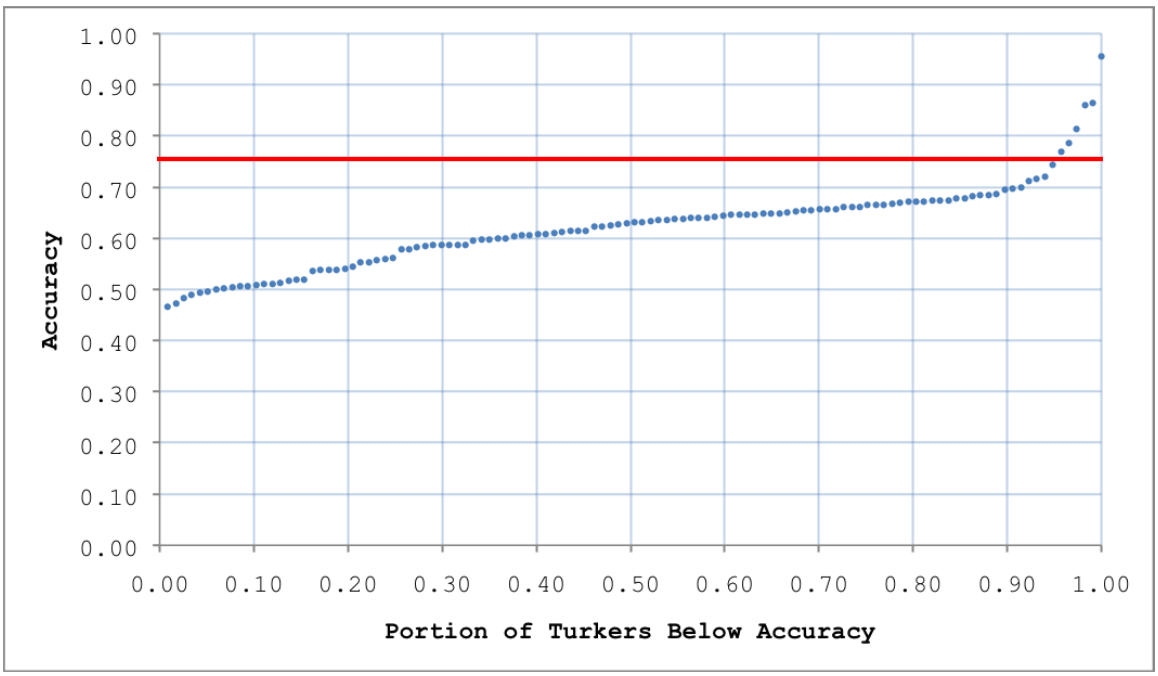

Figure 11: Human accuracy in rank order (100 responses or more), with classifier performance (line)

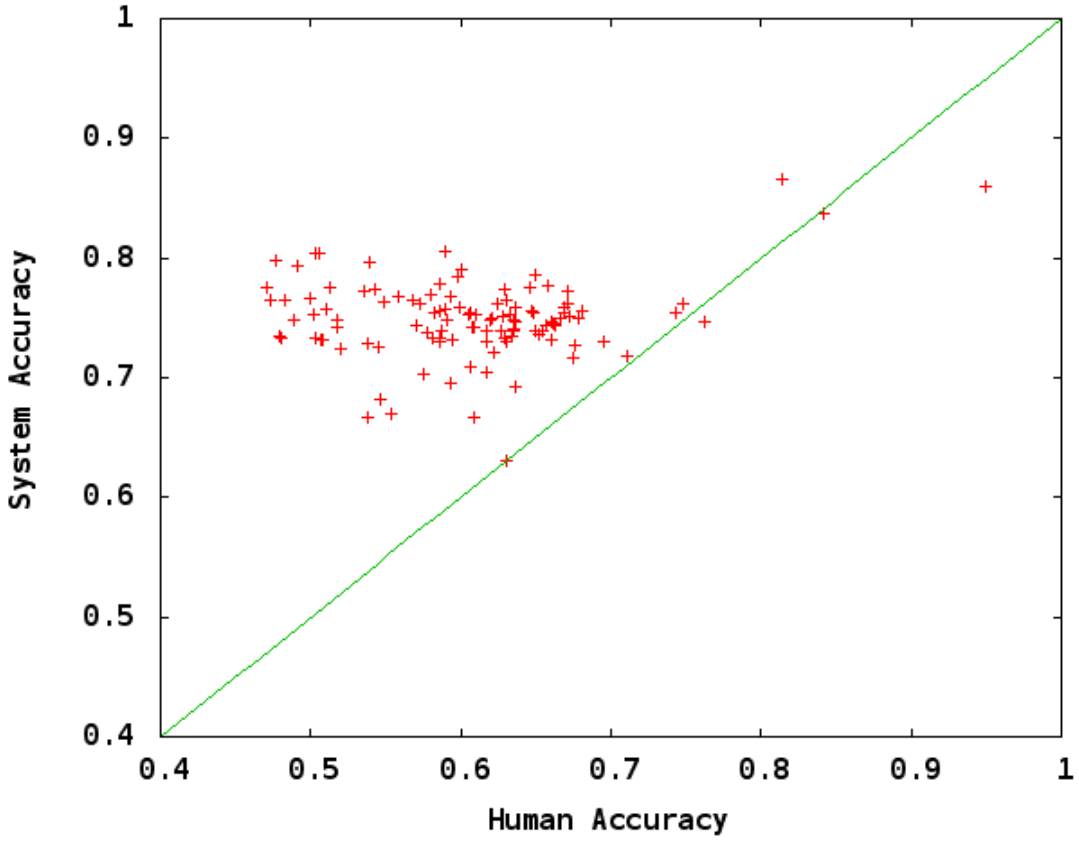

Figure 12: Classifier vs. human accuracy on the same subsets (100 responses or more)

6 Conclusion

In this paper, we have presented several configurations of a language-independent classifier for predicting the gender of Twitter users. The large dataset used for construction and evaluation of these classifiers was drawn from Twitter users who also completed blog profile pages.

These classifiers were tested on the largest set of gender-tagged tweets to date that we are aware of. The best classifier performed at 92% accuracy, and the classifier relying only on tweet texts performed at 76% accuracy. Human performance was assessed on this latter condition, and only 5% of 130 humans performed 100 or more classifications with higher accuracy than this machine.

In future work, we will explore how well such models carry over to gender identification in other informal online genres such as chat and forum comments. Furthermore, we have been able to assign demographic features beside gender, including age and location, to our Twitter dataset. We have begun to build classifiers for these features as well.

Acknowledgements

The authors would like to thank the anonymous reviewers. This work was funded under the MITRE Innovation Program.

References

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2007. Mining the blogosphere: Age, gender, and the varieties of self-expression. *First Monday*, 12(9), September.

John D. Burger and John C. Henderson. 2006. An exploration of observable features related to blogger age. In *Computational Approaches to Analyzing Weblogs: Papers from the 2006 AAAI Spring Symposium*. AAAI Press.

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10. Association for Computational Linguistics.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

A.P. Dawid and A.M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1).

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Conference on Empirical Methods on Natural Language Processing*.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1).

Bill Heil and Mikolaj Jan Piskorski. 2009. New Twitter research: Men follow men and nobody tweets. *Harvard Business Review*, June 1.

Susan C. Herring, Inna Kouper, Lois Ann Scheidt, and Elijah L. Wright. 2004. Women and children last: The discursive construction of weblogs. In L. Gurak, S. Antonijevic, L. Johnson, C. Ratliff, and J. Reyman, editors, *Into the Blogosphere: Rhetoric, Community, and Culture of Weblogs*. `http://blog.lib.umn.edu/blogosphere/`.

David Huffaker. 2004. Gender similarities and differences in online identity and language use among teenage bloggers. Master's thesis, Georgetown University. `http://cct.georgetown.edu/thesis/DavidHuffaker.pdf`.

Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on Amazon Mechanical Turk. In *Proceedings of the Second Human Computation Workshop (KDD-HCOMP 2010)*.

Nick Littlestone. 1988. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2, April.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

Claire Cain Miller. 2010. Why Twitter's C.E.O. demoted himself. *New York Times*, October 30. `http://www.nytimes.com/2010/10/31/technology/31ev.html`.

Arjun Mukherjee and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, MA, October. Association for Computational Linguistics.

Sasa Petrovic, Miles Osborne, and Victor Lavrenko. 2010. The Edinburgh Twitter corpus. In *Computational Linguistics in a World of Social Media*. AAAI Press. Workshop at NAACL.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. In *2nd International Workshop on Search and Mining User-Generated Content*. ACM.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. 2006. Effects of age and gender on blogging. In *Computational Approaches to Analyzing Weblogs: Papers from the 2006 AAAI Spring Symposium*. AAAI Press, March.

Robin Wauters. 2010. Only 50% of Twitter messages are in English, study says. *TechCrunch*, February 1. `http://techcrunch.com/2010/02/24/twitter-languages/`.