Cross-domain gender detection in Twitter*

Mohsen Sayyadiharikandeh[1**], Giovanni Luca Ciampaglia[2], and Alessandro Flammini[1,2]

[1] Indiana University, School of Informatics and Computing, Bloomington IN 47408, USA,
[2] Indiana University, Network Science Institute, Bloomington IN 47409, USA

Abstract. Unstructured textual data from online profiles is often used in conjunction with other user metadata to mine, in a supervised fashion, the latent demographic attributes of social media users (e.g. age, gender, occupation). Supervised methods, however, require labeled training data, which are often expensive to generate, and thus it would be attractive to re-use models across different domains and groups, i.e. training on a labeled dataset in order to mine the same latent attributes in those datasets for which training labels are missing. However, online conversations are often influenced by a myriad of topics and other factors, such as external events, and thus not all the features generated from this kind of data may perform well in a cross-domain setting. Here we study which of the features commonly found in public user profiles are portable across domains. As benchmark we focus on the very common task of detecting the gender of Twitter users from their public profile information — tweets, screen name, and profile picture. Our approach, based on a boosted stacked classifier, outperforms the state of the art in the task. Using data from two very different samples of Twitter users — one drawn from the public random stream and one about a recent social movement — we show that screen name and profile picture generalize across domains well, while text does not. Social media platforms have become attractive sources of data for computational approaches to social modeling, mainly due to their rapid growth and for the surprising ability to offer insight into real-world phenomena. Cross-domain user mining methods can help computational social science research by providing a richer and more accurate context to social phenomena.

1 Introduction

Computational approaches to social modeling increasingly rely on data from online social media, which are attractive in part for the vast amount of structured and unstructured data that they generate. While valuable, social media data

* Submitted to CHASM'16. Please do not circulate without permission from the authors.

** Corresponding authors: msayyadi@umail.iu.edu

often lack proper demographic information that is often needed to give a proper context to the phenomenon under study.

The growing use of online social media makes indeed these platforms great resources for studies in a wide range of disciplines, including the social, behavioral, and economic sciences [31]. Nowadays social media host discussions about any imaginable topic, offering a wealth of potentially helpful data to inform studies of society [16], although a number of methodological challenges need to be taken into account. A major setback that limits the usefulness of data extracted from online social media is the fact we often ignore any demographic information about the authors of the messages one is taking into consideration for a specific study. For example, it would be hard to formulate predictions about the result of an upcoming political election using the opinions (assuming reliably) collected on Twitter if one ignores which segment of the general populations has expressed such opinions.

It would be therefore highly desirable to avail of methods to infer latent user features like gender, age, and other demographic information which are not usually asked upon registration on most of these platforms, or that may not available to researchers due to obvious privacy reasons [38,40].

Latent attribute inference, the computational discovery of "hidden" attributes, has become a topic of significant interest to both social media researchers and industry practitioners interested in social media mining activities like recommendation and personalization. Recent work on latent attribute inference, particularly on Twitter, includes efforts to detect gender [19], age [24], education, and political affiliations [6,30], to cite a few.

Here we propose a supervised learning approach to infer the gender of a Twitter user on the basis of the public information available in her account. A major challenge in using a user's posts to infer her gender is the fact that tweets are short and usually contain Internet slang. This — compared to other longer texts like, e.g. blog posts — makes linguistic analysis harder.

Also, content duplication is much more common in Twitter compared to other online platforms, due to the 'retweet' feature. Under the assumption that retweets do not carry information about the gender of the user who shares them, this typically leaves one with an effective dataset that is smaller compared to platforms with less or no duplication, making the prediction task harder.

In this paper we propose a new gender detection method for micro-blogging social media which leverages the tweets posted by users, their self-reported name (*screen name*), and their profile pictures (*profile avatar*). Our work is motivated by the following considerations:

- Text-based gender identification often requires rich textual features, which may be available only for a small portion of the social media users.
- Although some success has been reported on specific datasets, it is unclear how easily the set of discriminative features can be ported from one dataset to another. For example, in prior work a quite specific dataset was considered and it was found that 'Bilbao' and 'Llorente' are discriminative terms for gender; the authors admit that these two terms are unlikely to transfer

to other datasets [8]. Textual features may not be portable and they are, therefore, usually selected and engineered according to the context [18].
- The issue of portable features in text mining is important in several areas, from cross domain classifications [22] to gender detection [5,18]. Good examples are applications in which the dataset is generated by crawling social networks starting from a query or a set of seed documents. For example, in sentiment analysis tasks, messages with a particular affective state could be used as seeds. The language of messages retrieved in this way will be biased toward this choice, and as a consequence textual features engineered from them cannot be used for other datasets with different seeds or queries. Using avatar image and screen name, which are both independent of the seeds, can help us build more general models for different domains. Prior work indicating that the accuracy of gender inference depends on tweets volume [3] substantiates this claim, since the model cannot perform well for users with low amount of tweets.
- Building the gold standard for training purposes for any sample of users is a time-consuming task. Using a model that has been previously trained on portable features is considerably more efficient.
- To the best of our knowledge computer vision algorithms have never been used before to infer the gender of Twitter users from profile pictures. Interestingly much previous work used profile avatars to build gold standard datasets [5,29], but none of them has used the power of computer vision classifiers to automatically capture faces from images and predict user gender based on this information. Using information in profile avatars we can improve accuracy of gender classifier.

The contributions of our present work include:

1. Using stacked classifiers in conjunction with boosting, we achieved a 96% accuracy, to be compared with about 86% of current state-of-the-art alternative approaches.
2. We demonstrate the usefulness of portable features such as screen name or profile picture, which are in principle independent from language use; as mentioned above, portable features could be useful in cross-domain classification tasks, or simply when one wants to infer the gender of a sample of users that have been collected starting from a possibly biased query.
3. We show that model based on portable features can perform well when content is scarce, or when tweets do not include the discriminating terms studied before [18,19].

2 Related Work

Several different methods have been proposed for inferring latent attributes of users of the Web, weblogs, and social media. A number of domain-specific tools for gender detection have been developed, e.g. in speech transcriptions [32], blogs [4,11], movie reviews [27], e-mail [10], and search queries [14,35].

Where social media are concerned, link-based and group-based classification has been proposed in Facebook to study how visible attributes like friendship and group membership can inform the inference of sensitive attributes like political views [38]. The accuracy of gender detection in Facebook was improved by focusing on the names associated with users [34]. In the context of predicting gender and age of FB users, the use of non-I.I.D. multi-instance learning has been proposed to prevent the learning algorithm produce a biased model [33].

Much work has also been done regarding inferring age, gender, ethnicity and political orientation specifically for Twitter users. One of the first works in this area identified gender, age, regional origin, and political orientation using stacked-SVM-based classification algorithms over a rich set of features [30]. The authors considered both the content of the tweets and the writing style to discriminate between male and female. It has been also shown that tweets volume affect the accuracy of the classifier [3]. Other work leverages, together with more traditional features, the set of celebrities followed for gender identification [20].

Prior work has tried to detect political orientation, ethnicity, and gender by leveraging user behavior, network structure, and the language of users, again on Twitter [29].

Most of the aforementioned approaches focus on English-speaking users. Research has focused on detecting gender in non-English locales using language-specific features [5]. Unsurprisingly, it has been found that models trained on data from one language cannot be used for other languages. In other words, detecting user attributes in non-English contexts shows the need for better *cross-domain classification* techniques and, by extension, of more portable features that are robust against different languages and locales.

Finally, one study took into account what is perhaps one the most indicative signal of the gender of a person: the first name [19]. The authors used this information along with textual features obtained from tweets, and attained an accuracy of 86%. This work is also valuable because — at least within the scope of a single language — the first name is a portable feature. However, in another work Ruths and Liu used the aforementioned model to infer the gender composition of commuter populations [18]; they found that the model was not general enough. This indicates that even though a feature may be useful within one context, it may not be portable to another, and thus that more than one portable feature (for example name *and* profile picture) may be needed.

It should also be mentioned that some obvious non-textual features, like network structure and communication behavior, have been already considered [30]. However, no satisfactory signal could be extracted from such features. Indeed, understanding the structure of gender interactions dates back to the seminal work of sociologists in the early 19th century [23].

Thus, finding a general and reliable inference model for gender detection is still an open problem, and the use of portable features seems a promising direction to explore. As stated in the literature the accuracy of the state of the art is generally between 80% and 86% [5]. An accuracy of 90% was reported in

other work [3], but it was relative to a set of users quite different from the typical anglophone Twitter user.

Some authors considered a small dataset collected using the Twitter REST API over a short period of time [8]. Feature selection resulted in good accuracy of the method. At the same time the authors acknowledge the specificity of the dataset: discriminative terms like 'Bilbao' or 'Welbeck' for men would most likely not be useful on other datasets.

In summary, the unstructured text posted by social media users is a powerful source of features that can be used for prediction, but such features are too often dataset dependent. Also, substantial feature engineering is needed in order to produce the rich structured datasets suitable for training the automated classifiers presented in those works.

Here we introduce a new feature: the image used by Twitter users in their profiles. This information, together with textual features and the screen name, let us build a richer dataset and attain higher accuracy than the current state of the art.

3 Proposed Method

As mentioned in prior work, when considered in isolation, neither screen names, language, or profile pictures are satisfactorily discriminative of the gender of a generic user.

Here, we use a stacked classifier approach which, by chaining multiple estimators, yields a more robust classifier. Figure 1 shows the proposed framework. The classifier is structured in two layers. In the first layer three base classifiers receive in input the sequence of tweets, the screen name, and the profile avatar of a user, respectively. Note screen name and url to profile avatar can be found in user's profile. For tweets we should crawl tweets authored by users and also download image for user using automatic tools in order to be able to use this framework. However, these base classifiers produce each a predicted gender label and, in the case of the text and image classifiers, a measure of confidence. This information is weighted into a meta classifier, which forms the last layer.

The approach described goes under the name of *stacked generalization*. Stacked generalization is a well-known ensemble approach, first studied by Wolpert [36]. The effectiveness of this approach was studied both in latent attribute inference in Twitter [30] and other domains [37,1].

Intuitively, training successfully such a model entails two outcomes: first, in the first layer each of the base classifiers should find the subset of observations it can predict optimally; second, in the second layer the meta-classifier should be able to combine the predictors from the first layer.

The benefit of this approach is that it allows to combine multiple weak classifiers from different sources, without having to retrain all of them at once. For comparison, one could imagine training a neural network that takes all the aforementioned inputs (i.e., picture, screen name, and textual tokens) and then learns a single model. If we want to add another input, we would need to retrain the

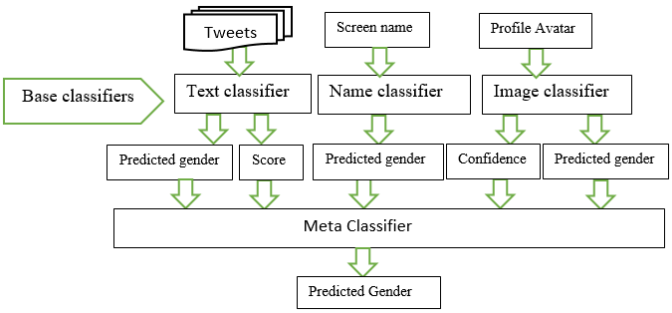

Fig. 1. Framework used for gender inference.

Table 1. Class composition of ground truth data.

	All	Female	Male
Users	12,681	8,232	4,449
Percentage		65%	35%

whole model, whereas with a stacked classifier one would simply need to add a base classifier and retrain the meta classifier. Note that ensemble learning generally works well when the predictors it aggregates are weakly correlated among each other.

3.1 Data

In order to evaluate our method and compare it with other work in the domain we decided to work with the dataset published by Ruths and Liu [19], which includes 12,680 Twitter users. For each user, the numeric ID and a binary gender label are provided.

The authors claim that this public dataset is representative of the general Twitter population [19], since they selected a representative sample of users who had posted at least 1,000 tweets over the lifespan of their accounts. Good performance on this data implies good chances the model will work well with arbitrary Twitter users.

As we can see from Table 1 the gender composition of the dataset is significantly skewed towards female users; this is compatible with Twitter being a platform more popular among women, according to the observation reported by prior research according to which 55% of Twitter users are female [21].

We used the Twitter REST API to download profile avatar and screen name for each user represented in the base dataset, and collected her tweets from a comprehensive historical archive of tweets obtained from the Twitter Garden-Hose, a 10% random sample of all tweets [7]. Retweets and other simple form of near duplicates were removed. To better gauge the importance of portable

Table 2. Training dataset for text classifier

Dataset	Date range	Tweets (avg)	σ
D1	Jan 2014-Dec 2015	63	148
D2	Jan 2010-Dec 2014	530	871

Table 3. Performance of the base classifier.

Classifier	Dataset	Acc.	Rec.	F-score	Coverage
Name	D1 + D2	88%	88%	88%	51%
Image	D1 + D2	87%	88%	88%	32%
Text	D1	74%	63%	68%	100%
Text	D2	82%	92%	86%	100%

features, we collected two datasets, D1 and D2, that differ by the number of tweets per user collected. Our prior work [3] suggests that more tweets increase accuracy. Table 2 shows basic statistics for the two datasets.

3.2 Name Classifier

According to prior research the information in the self-reported screen name of Twitter users can be exploited for gender inference [19]. We employ the screen name as one of our features. To simplify the task of data extraction of such feature, we used the Microsoft Discussion Graph Tool (DGT) [15].

Some Twitter users use their real first name as screen name. Many may instead choose nicknames or names that are generally less gender-revealing. DGT generates the label 'unknown' when it is not able to classify a user with confidence. We define a notion of coverage, that is, the fraction of cases for which DGT emits a label other than 'unknown'. In our sample, the coverage was 51%. If we consider only this cases, the accuracy of DGT is 87.89%, meaning that many screen names are highly gendered, and thus descriptive enough to be used for inference for that subset of the users. Table 3 gives a summary of the results of the DGT-based name classifier.

3.3 Image Classifier

There has been much progress in the area of computer vision on methods for acquiring, processing, analyzing, and understanding images with the adoption of deep learning [17]. As mentioned earlier, and to the best of our knowledge, exploiting social media profile avatars has not been given much attention in the gender detection literature, at least compared to inference methods based on text and behavioral and structural data. Interestingly, photos have been used

before to build gold standard datasets for these tasks [5], which shows implicitly the potential importance of profile avatars.

In prior work a sample of 15,000 random users was drawn and the profile avatars of these users were considered manually [29]. The authors reported that 57% of user profile pictures reflect the gender of their users. On the other hand, 20% of profile avatars were depicting a celebrity, or people other than the user.

In our project we use Face++, a naive deep learning face recognition tool [39]. Face++ uses deep convolutional networks and is trained on a large dataset collected on the Web. It achieves state-of-the-art performance on the LFW benchmark dataset [12] with 99.5% recognition accuracy, surpassing human raters on the same dataset [39].

Social media profile pictures pose several challenges for gender detection: first, several profiles show only the default picture; second, some users use the pictures of their pets or other subjects that not convey information for predicting gender, as already discovered before [29]; third, there are cases in which the profile picture displays more than one face in the photo e.g. the user among a group of friends; fourth, any human faces present in the picture may be too small for the detection to work reliably. In the case in which Face++ detects more than one face, we go for the gender of the majority, and break ties in favor of 'female' which, as already mentioned, is reported to be majority group on Twitter [21].

Face++ computes a confidence score along with the predicted gender, indicating how much the algorithm is confident about the prediction. We take this confidence value as an input feature for the meta classifier, since it can help us find the subset of users for whom the profile photo is indicative of their gender.

We experimented the task of gender detection using only Face++ on the dataset and we got 87.48% accuracy with 32% coverage. The average confidence of prediction for high-confidence samples was 90.3%, while for the others was 81%. We randomly selected 100 wrong and right predictions and performed a t-test and compared the average confidence achieved by Face++. The average confidence for the "wrong prediction" set was lower than that for the "right prediction", and the difference was statistically significant. Table 3 reports the results of the image classifier on the full dataset.

3.4 Text classifier

As a preprocessing step we removed any stop word and transformed tweets into vectors of unigrams. More aggressive forms of pre-processing, like spell checking and translation of slang terms into regular English, did not enhance the accuracy of the classifier, and so we decided not to use them. The resulting dictionary is approximately 5×10^4 and 5.2×10^4 terms big for dataset D1 and D2, respectively. We fed the sparse vectors to SVMLight [13] and tested different kernel transformation functions, finding that linear kernels were the fastest and most accurate choice.

Table 3 shows the performance of the best classifier trained with SVMLight on both D1 and D2 datasets. Note that in the case of this classifier there is not

Table 4. Performance of the boosted stacked classifier.

Dataset	Acc.	Rec.	F-score
D1	87.1%	88.4%	87.7%
D2	95.9%	97.1%	96.5%

an obvious measure of coverage to use. Looking at incorrect predictions from this classifier in dataset D2, we found that the average number of tweets for users in this category was 273 with a standard deviation of $1,078.42$. We also selected users with less than 30 tweets from dataset D2 and observed that about half of them (54.55%) were correctly classified. In the next section, we compare the performance of the meta classifier on the same dataset to get the idea of the improvements gained from stacking all classifiers.

3.5 Stacked classifier via boosting

The predictions of DGT, Face++, and our SVM-based text classifier, along with the confidence score from Face++ and the probabilistic score of SVM form a set of features that we feed into the meta classifier, see Figure 1.

When it is difficult to find a single, highly accurate prediction rule, boosting algorithms like AdaBoost can combine a number of weak classifiers into a more powerful one. AdaBoost uses an optimally weighted majority vote of weak classifiers. This approach is mostly effective in those cases that have been misclassified the most by the underlying weak classifiers.

We also experimented with a simpler logistic regression classifier as meta classifier, attaining results that are competitive with the state of the art. This implies that most of the benefits from stacking derives from using a rich set of features like the one described above. In the following, however, we only report the results obtained via boosting.

4 Results

We used Python as the main language for processing the data, and in particular relied on Scikit-learn [28] to develop the meta classifier described in the previous section. Experiments were performed on a personal computer with 4GB of RAM and Intel© Core™ i5 CPU.

Table 4 shows precision, accuracy, recall, and F1-score of the meta classifier, computed using 5-fold cross-validation. Training set is created by choosing 80 percent of dataset randomly and the rest is considered as a test set in both D1 and D2. In Figure 2 we compare these results with the accuracy reported by Ruths and Liu [19] on the same sample of users. As a baseline for both methods, we show the performance of the maximum likelihood estimator, i.e. the classifier

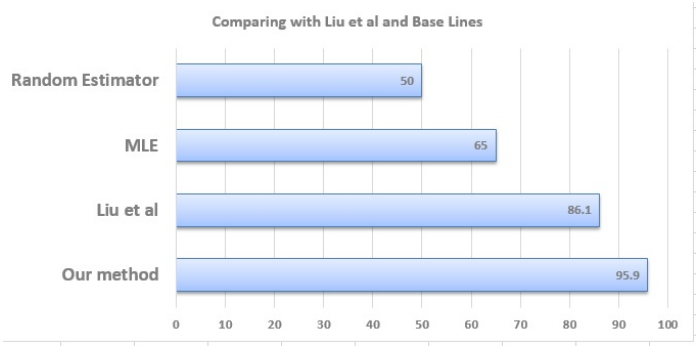

Fig. 2. Performance of the stacked classifier; comparison with random, base rate from the data, and prior art [19].

Table 5. Gender retrieval task; performance of name and image stacked classifier.

Gender	Acc.	Rec.	F-score
Female	77.7%	95.1%	85.5%
Male	88.7%	58.9%	70.8%

obtained by choosing a label with probability proportional to the fraction of female users in the sample.

Focusing on wrong predictions, we compute the rate of false positives, which is the fraction of male users out of all users who were misclassified (i.e. 'Female' classified as 'Male' *plus* 'Male' classified as 'Female'). This fraction is equal to 84%, meaning that the overwhelming majority of wrong predictions happens for male users.

To better understand why this is the case, we can reframe the problem of gender detection as a retrieval task. So far we have indeed treated the problem as a binary classification task, assuming that 'Female' was the 'True' label, and 'Male' the 'False' one. We can instead consider each gender as the set of "relevant" examples, and compute accuracy and recall scores accordingly. Using dataset D2, this exercise yields two sets of scores — one per gender — which are shown in Table 5. In particular, we see that the recall for 'Male' (58.9%) is much lower than that for 'Female' (95.1%), showing that retrieving 'Male' labels is much harder than retrieving the 'Female' ones.

Because neither DGT nor Face++ had problems in detecting individual gender when used in isolation (see Table 3), this indicates that males are more 'hidden' in terms of using self-reported names and profile photos than females. Interestingly, this result was already reported in prior work [3].

As noted before, ensemble methods like boosting work well when the correlation between its inputs is not high. As a form of diagnostics, Table 6 shows the

Table 6. Correlations among the base classifiers.

	Pearson	**Spearman**
Image Name	0.27	0.45
Image Text	0.34	0.56
Name Text	0.42	0.58

Table 7. BLM dataset.

Dataset	**Tweets (avg)**	σ
BLM1	17.8	63.9
BLM2	789.1	1549.4

Pearson's and Spearman's correlation coefficients of the labels emitted by each of the three base classifiers. Dataset D2 was used to estimate the coefficients. To compute these correlations, the 'Male' label was mapped to -1, the 'Female' one to $+1$, and the 'Unknown' label, which is returned by both DGT and Face++, to 0. Pairwise correlations range from moderate to medium.

To assess whether the stacked classifier is suitable for use in a cross-domain classification context, we took advantage of our two datasets; we trained the model on D1 (resp. D2) and tested it on D2 (resp. D1). This approach simulates two different cross-domain classification scenarios. The first case, in which we train on the smaller dataset and test on the larger one, simulates the case in which the model must be able to overcome the bias induced by a small dataset. The other one (training on large dataset, testing on small) should simulate the more interesting case in which a model with portable features is applied to a different domain. In the former the stacked classifier attains an accuracy of 95.5%, while in the latter an accuracy of 83.4%. As a term of comparison, for the same two tasks the text classifier attained an accuracy of 80.5% and 68.0%, respectively.

We also tested the performance of our stacked classifier for cross domain classification task on a different dataset, obtained for the study of #BlackLivesMatter, a prominent US social movement [26]. Table 7 shows some description of two dataset derived from BLM dataset.

We evaluated the capability of a stacked classifier for cross domain classification task with different feature sets on BLM dataset. Table 8 shows the accuracy on both BLM1 and BLM2 dataset using different features with 5 fold cross validation. Our results indicates that name and profile avatar are portable features, while text cannot be relied for cross domain classification task. Using our stacked classifier with inter domain classification we got 93.4 percent in accuracy which shows applicability of our method in different datasets.

Table 8. Performance of stacked classifier on BLM dataset.

Dataset	feature set	type	Acc.
BLM1	text	Inter.	58.1%
BLM1	text	Cross.	58.9%
BLM1	text + face	Inter.	75.3%
BLM1	text + face	Cross.	63.3%
BLM1	text + name	Inter.	78%
BLM1	text + name	Cross.	67.8%
BLM1	face + name	Inter.	76%
BLM1	face + name	Cross.	76%
BLM1	text + face + name	Inter.	85%
BLM1	text + face + name	Cross.	72.8%
BLM2	text	Inter.	71.9%
BLM2	text	Cross.	59.4%
BLM2	text + face	Inter.	88.3%
BLM2	text + face	Cross.	62.6%
BLM2	text + name	Inter.	89.6%
BLM2	text + name	Cross.	63.5%
BLM2	face + name	Inter.	76%
BLM2	face + name	Cross.	76%
BLM2	text + face + name	Inter.	93.4%
BLM2	text + face + name	Cross.	71.1%

We also focused on users who had less than 30 tweets recorded in dataset D2, and obtain an accuracy of 90.9%. This shows the generality of our model and how employing portable features like screen name and profile avatar can enhance the performance of classifier, even in situations where text is scarce.

To understand the relative importance of each component of the meta classifier, we also tested a partial stacking approach, i.e. using only a subset of the input features to train and test AdaBoost. Table 9 shows the results of this exercise on dataset D2, using the three possible pairwise combinations of the three base classifiers.

Finally, Figure 3 shows the ROC curve — the true positive rate plotted against the false positive rate. The area under the ROC curve is 0.97.

4.1 Limitations

We manually checked a random sample of the users that were misclassified (67% male). These users had on average less tweets (271, $\sigma = 265.7$) than the broader sample. Face++ failed to give a label in 83% of these cases. This is compatible with the claims about its low false positive rate [39,29]. Regarding the screen names of these cases, DGT produced an 'Unknown' label 61% of the times.

Table 9. Performance of partial stacking.

Classifier	Acc.	Rec.	F-score
MLE	65.0%	65.0%	65.0%
image + name	79.1%	77.0%	78.0%
image + text	89.9%	89.3%	89.6%
name + text	88.7%	89.1%	88.9%

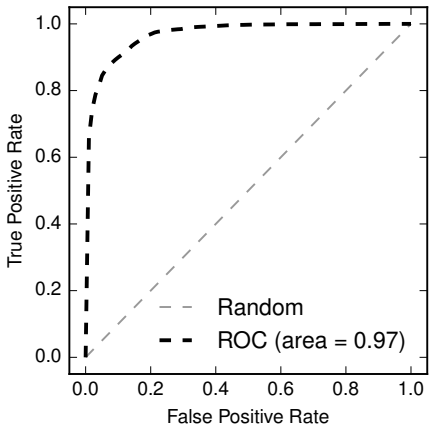

Fig. 3. ROC Diagram of the stacked classifier.

In some instances, some of the base classifiers were in disagreement with each other. For example, in some cases Face++ gave a correct prediction with high confidence, and our SVM text classifier an incorrect prediction with a low score. Nonetheless, AdaBoost was still able to give the correct answer. In contrast, the simpler logistic regression meta classifier gave more weight to text and produced an incorrect prediction.

5 Conclusion

In this paper we used data from Twitter to study the task of gender detection in social media. The main contribution of our method was to employ a stacked classification framework, and to include the output of face recognition algorithm to improve the overall accuracy. To the best of our knowledge, this is the first time that computer vision algorithms are used for inferring the gender of Twitter users.

Our work can motivate the research community to bridge the gap between computer vision and text mining algorithms and paves the way for applica-

tion of these techniques to studies of gender-related phenomena, such as harassment [2,9] or gender gap [25]. We envision our approach to be particularly useful in cases where the amount of text for each user is limited, or when the users are collected starting from a seed query that may substantially bias the language used in the tweets and also our trained model can also be used for finding gender of new unlabeled datasets using cross-domain classification.

We tested the hypothesis of combining different gender predictors together, to create a single more accurate estimator for gender prediction with a lower error rate. Individual classifiers of text, profile avatar, and screen name were employed in order to build a rich vector of features. Our experimental results show that the fusion of these three classifiers overcomes their individual shortcomings, yielding a better overall classifier. As future works, we plan to apply our framework to other platforms, like Google+, and experiment with alternative classifiers for text.

Acknowledgment

The authors would like to thank Alexandra Olteanu, Derek Ruths, and Wendy Liu for making their respective dataset available. This work was partially supported by the Indiana University Network Science Institute.

References

1. Álvarez, A., Sierra, B., Arruti, A., López-Gil, J.M., Garay-Vitoria, N.: Classifier subset selection for the stacked generalization method applied to emotion recognition in speech. Sensors 16(1), 21 (2016), http://www.mdpi.com/1424-8220/16/1/21
2. Bartlett, J., Norrie, R., Patel, S., Rumpel, R., Wibberley, S.: Misogyny on twitter (2014), http://www.demos.co.uk/files/MISOGYNY_ON_TWITTER.pdf, last accessed: 2016-05-16
3. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1301–1309. EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2145432.2145568
4. Burger, J.D., Henderson, J.C.: An exploration of observable features related to blogger age. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. pp. 15–20 (2006)
5. Ciot, M., Sonderegger, M., Ruths, D.: Gender Inference of Twitter Users in Non-English Contexts. In: Proceedings of EMNLP (2013)
6. Conover, M., Gonçalves, B., Ratkiewicz, J., Flammini, A., Menczer, F.: Predicting the political alignment of Twitter users. In: Proc. 3rd IEEE Conference on Social Computing (SocialCom) (2011)
7. Davis, C.A., Ciampaglia, G.L., Aiello, L.M., Chung, K., Conover, M.D., Ferrara, E., Flammini, A., Fox, G.C., Gao, X., Gonçalves, B., Grabowicz, P.A., Hong, K., Hui, P.M., McCaulay, S., McKelvey, K., Meiss, M.R., Patil, S., Peli Kankanamalage, C., Pentchev, V., Qiu, J., Ratkiewicz, J., Rudnick, A., Serrette, B., Shiralkar, P., Varol,

O., Weng, L., Wu, T.L., Younge, A.J., Menczer, F.: OSoMe: The IUNI observatory on social media. PeerJ Preprints 4, e2008v1 (2016)
8. Deitrick, W., Miller, Z., Valyou, B., Dickinson, B., Munson, T., Hu, W.: Gender identification on twitter using the modified balanced winnow. Communications and Network Vol.04No.03, 7 (2012), http://www.scirp.org/journal/PaperInformation.aspx?PaperID=22061
9. Fulper, R., Ciampaglia, G.L., Ferrara, E., Menczer, F., Ahn, Y., Flammini, A., Lewis, B., Rowe, K.: Misogynistic Language on Twitter and Sexual Violence. In: Proc. ACM Web Science Workshop on Computational Approaches to Social Modeling (ChASM) (2014), http://dx.doi.org/10.6084/m9.figshare.1291081
10. Garera, N., Yarowsky, D.: Modeling latent biographic attributes in conversational genres. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2. pp. 710–718. ACL '09, Association for Computational Linguistics, Stroudsburg, PA, USA (2009), http://dl.acm.org/citation.cfm?id=1690219.1690245
11. Herring, S.C., Paolillo, J.C.: Gender and genre variation in weblogs. Journal of Sociolinguistics 10(4), 439–459 (2006), http://dx.doi.org/10.1111/j.1467-9841.2006.00287.x
12. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep., Technical Report 07-49, University of Massachusetts, Amherst (2007)
13. Joachims, T.: Making large-scale svm learning practical. LS8-Report 24, Universität Dortmund, LS VIII-Report (1998)
14. Jones, R., Kumar, R., Pang, B., Tomkins, A.: "i know what you did last summer": Query logs and user privacy. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management. pp. 909–914. CIKM '07, ACM, New York, NY, USA (2007), http://doi.acm.org/10.1145/1321440.1321573
15. Kıcıman, E., Counts, S., Gamon, M., De Choudhury, M., Thiesson, B.: Discussion graphs: Putting social media analysis in context. In: Intl. Conf. on Weblogs and Social Media (ICWSM-14). AAAI (June 2014), http://research.microsoft.com/apps/pubs/default.aspx?id=210256
16. Lazer, D., Pentland, A.S., Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al.: Life in the network: the coming age of computational social science. Science 323(5915), 721 (2009)
17. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436–444 (May 2015), http://dx.doi.org/10.1038/nature14539
18. Liu, W., Al Zamal, F., Ruths, D.: Using social media to infer gender composition of commuter populations. Sixth International AAAI Conference on Weblogs and Social Media abs/1405.6667 (2012)
19. Liu, W., Ruths, D.: what's in a name? using first names as features for gender inference in twitter. AAAI Spring Symposium Series abs/1405.6667 (2013)
20. Ludu, P.S.: Inferring gender of a twitter user using celebrities it follows. CoRR abs/1405.6667 (2014), http://arxiv.org/abs/1405.6667
21. Mislove, A., Lehmann, S., Ahn, Y.Y., Onnela, J.P., Rosenquist, J.N.: Understanding the Demographics of Twitter Users. In: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11). Barcelona, Spain (July 2011)

22. Mohammad, S.: Portable features for classifying emotional text. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 587–591. NAACL HLT '12, Association for Computational Linguistics, Stroudsburg, PA, USA (2012), http://dl.acm.org/citation.cfm?id=2382029.2382123
23. Moreno, J.L.: Who shall survive?: A new approach to the problem of human interrelations. Nervous and Mental Disease Publishing Co, Washington, DC, US (1934)
24. Nguyen, D.P., Gravel, R., Trieschnigg, R.B., Meder, T.: "how old do you think i am?" a study of language and age in twitter. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, MA, USA. pp. 439–448. AAAI Press, Palo Alto, CA, USA (July 2013)
25. Nilizadeh, S., Groggel, A., Lista, P., Das, S., Ahn, Y.Y., Kapadia, A., Rojas, F.: Twitter's glass ceiling: The effect of perceived gender on online visibility. In: Tenth International AAAI Conference on Web and Social Media (2016)
26. Olteanu, A., Weber, I., Gatica-Perez, D.: Characterizing the demographics behind the #BlackLivesMatter movement. In: Proceedings of the AAAI Spring Symposia on Observational Studies through Social Media and Other Human-Generated Content (SSS'16 OSSM). Stanford, US (2016)
27. Otterbacher, J.: Inferring gender of movie reviewers: Exploiting writing style, content and metadata. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. pp. 369–378. CIKM '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1871437.1871487
28. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
29. Pennacchiotti, M., Popescu, A.M.: A machine learning approach to twitter user classification. In: Proceedings of Fifth International AAAI Conference on Weblogs and Social Media. ICWSM11 (2011)
30. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in twitter. In: Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents. pp. 37–44. SMUC '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1871985.1871993
31. Ruths, D., Pfeffer, J.: Social media for large studies of behavior. Science 346(6213), 1063–1064 (2014)
32. Singh, S.: A pilot study on gender differences in conversational speech on lexical richness measures. Sixth International AAAI Conference on Weblogs and Social Media abs/1405.6667 (2012)
33. Song, H.J., Son, J.W., Park, S.B.: Identifying user attributes through non-i.i.d. multi-instance learning. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 1467–1468. ASONAM '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2492517.2492597
34. Tang, C., Ross, K., Saxena, N., Chen, R.: What's in a name: A study of names, gender inference, and gender behavior in facebook. In: Proceedings of the 16th International Conference on Database Systems for Advanced Applications. pp. 344–356. DASFAA'11, Springer-Verlag, Berlin, Heidelberg (2011), http://dl.acm.org/citation.cfm?id=1996686.1996731
35. Weber, I., Castillo, C.: The demographics of web search. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in In-

formation Retrieval. pp. 523–530. SIGIR '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1835449.1835537
36. Wolpert, D.H.: Stacked generalization. Neural Networks 5, 241–259 (1992)
37. Wu, C.H., Liang, W.B.: Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. IEEE Trans. Affect. Comput. 2(1), 10–21 (Jan 2011), http://dx.doi.org/10.1109/T-AFFC.2010.16
38. Zheleva, E., Getoor, L.: To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles. In: Proceedings of the 18th International Conference on World Wide Web. pp. 531–540. WWW '09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1526709.1526781
39. Zhou, E., Cao, Z., Yin, Q.: Naive-deep face recognition: Touching the limit of LFW benchmark or not? CoRR abs/1501.04690 (2015), http://arxiv.org/abs/1501.04690
40. Zimmer, M.: "but the data is already public": on the ethics of research in facebook. Ethics and Information Technology 12(4), 313–325 (2010), http://dx.doi.org/10.1007/s10676-010-9227-5