Empirical Evaluation of Profile Characteristics for Gender Classification on Twitter

Jalal S. Alowibdi[1, 2], Ugo A. Buy[1] and Philip Yu[1, 2]
[1]Department of Computer Science
University of Illinois at Chicago
[2]Faculty of Computing and Information Technology
King Abdulaziz University
{jalowibd,buy,psyu}@cs.uic.edu

Abstract—Online Social Networks (OSNs) provide reliable communication among users from different countries. The volume of texts generated by OSNs is huge and highly informative. Gender classification can serve commercial organizations for advertising, law enforcement for legal investigation, and others for social reasons. Here we explore profile characteristics for gender classification on Twitter. Unlike existing approaches to gender classification that depend heavily on posted text such as tweets, here we study the relative strengths of different characteristics extracted from Twitter profiles (e.g., first name and background color in a user's profile page). Our goal is to evaluate profile characteristics with respect to their predictive accuracy and computational complexity. In addition, we provide a novel technique to reduce the number of features of text-based profile characteristics from the order of millions to a few thousands and, in some cases, to only 40 features. We prove the validity of our approach by examining different classifiers over a large dataset of Twitter profiles.

***Keywords*- Color-based features, profile characteristics, phonemes as features, color quantization, social networks, language independence.**

I. Introduction

Online Social Networks (OSNs) have grown at a stunning rate over the past decade. They are now a part of the lives of dozens of millions of people. The growth in the user base has led to a dramatic increase in the volume of generated data. The onset of OSNs has stretched the traditional notion of "community" to include groups of people who have never met in person but communicate with each other through OSNs to share knowledge, opinions, interests and activities.

Our long-term goal is gender identification in online social networks with an emphasis on accuracy, computational efficiency and scalability of gender predictions. We are especially interested in language-independent methods. Only around 50% of Twitter messages are in English [1]. Our Twitter dataset alone contains 34 different languages. An estimate breakdown of language use in our dataset shows that around 69% of users are English speaking with the remaining 31% distributed over 33 languages.

Here we explore gender identification using only user profiles. Our approach is based on three characteristics for each user profile: (1) first name, (2) user name, and (3) profile colors. Profile colors include the background color, text color, link color, sidebar fill color and sidebar border color. We conducted extensive empirical studies on a large dataset of Twitter users in order to assess the relative strengths and weakness of these characteristics.

To date most existing approaches to gender classification on Twitter depend heavily on an analysis of text in posted messages, aptly called tweets; however, the strength of the above three profile characteristics is currently unknown. Burger et al. use four different characteristics from a user's profile and posts (i.e., first name, user name, description and tweets) for gender classification [2]. Alowibdi et al. use the five color features from a user's profile (i.e. background color, text color, link color, sidebar fill color, sidebar border color) [3]. Liu and Ruths use only first names for gender classification [4]. Other works for gender classification use only user posts in order to identify gender [5], [6], [7]. Except for our method [3], all existing approaches to gender classification on Twitter use word-based n-grams resulting in a huge feature space consisting of unique words and word combinations extracted from tweets. The size of the resulting feature sets is often in the order of many million features [2].

Our work is different from existing methods because of its simplicity and the range of profile characteristics that we consider. We defined a phoneme-based technique for reducing the number of features. Our method typically results in a reduction in feature space size by two to four orders of magnitude.

In the sequel we report our empirical results on different profile characteristics for gender classification. In particular, we predict automatically the gender value of users based on their profile preferences. We analyzed user profiles with different classifiers in the Konstanz Information Miner (KNIME), which uses the Waikato Environment for Knowledge Analysis (WEKA) machine learning package [8], [9]. Our main contributions are outlined below.

- We define a new phoneme technique for predicting gender, which sharply reduces feature set size to a few thousands features at the most, and in some cases only 40 features, from several million features.
- We compared empirically different profile characteristics in order to find the most accurate gender indicators.
- We validated our approach by analyzing different classifiers over a large dataset of Twitter profiles. Our results show that each profile characteristic can provide reasonably accurate gender predictions.

The remainder of this paper is organized as follows. In Section 2, we briefly summarize related work on gender clas-

sification. In Section 3, we detail our proposed approach. In Section 4, we report our empirical results from different classifiers and we analyze these results.

II. Related work

Many authors have investigated gender classification by using text sentiment in blogs, articles, and forum platforms [10], [11], [12], [13], [14], [15], [16], [17], [18]. Those authors explored a variety of methods, including word-frequencies, writing styles, Part-Of-Speech (POS) n-gram, POS tags, unigrams, word frequencies, word classes, POS patterns, POS contents and POS style metrics to analyze text. In general, all existing approaches depend heavily on posted text.

Burger et al. [2], Alowibdi et al. [3], Liu and Ruths [4], Alzamal et al. [5], Liu et al. [6] and Rao et al. [7], worked on gender classification in Twitter. In particular, Burger et al. [2] apply the n-gram feature model to four different characteristics, that is, name, user name, description and tweets. Their method results in some 15 million features. We used a color-based feature model that takes advantage of five different characteristics of profile layout colors (i.e., background color, text color, link color, sidebar fill color, sidebar border color) resulting in about 500 features [3]. Liu and Ruths [4] performed gender prediction using a first name feature-based approach. Alzamal et al. [5] and Liu et al. [6] applied the n-gram feature model to about 400 profiles and their tweets. Rao et al. [7] employ the sociolinguistic-feature model, n-gram feature model and stacked model to analyze text sentiment in posted tweets. They have about 1.2 million features. In general, approaches based on text analysis generate in the order of millions of features [3].

In summary, most existing researchers explored gender classification by applying various methods for text analysis either to tweets or posted text in blogs. Admittedly, the obstacles to using text sentiment are high computational complexity (resulting from millions of features) and language dependency. We rely on profile colors and a phonics-based n-gram transformation to dramatically reduce complexity while retaining a high degree of accuracy.

III. Proposed approach

Our approach can be summarized as follows:

1. We harvested a large dataset of Twitter profiles.
2. We identified the "ground truth" of a user's gender by following the links from the profiles to other OSNs.
3. We applied the Google Input Tools (GIT) to convert the characters of different languages to characters in English language.
4. We converted first names and user names to phoneme sequences.
5. We trained, tested and validated our gender predictions using different classifiers.

In Step 1 we harvest names, username, background colors, text colors, link colors, sidebar fill color and sidebar border colors from user profiles. For color-based features, we apply a color quantization and sorting procedure (i.e., normalization [3]) to reduce the number of colors from around 17 million unique colors (features) to only 512 unique colors.

Twitter profiles do not include an explicit gender field. Thus, in Step 2 we identify Twitter profiles with an external link to another profile (e.g., a Facebook profile) for the same user. If the other profile includes an explicit gender declaration, we use that declaration as the ground truth for the gender of that user.

In Step 3, we convert the alphabet of different languages than English (e.g., Japanese, Chinese, and Arabic) to characters in English with GIT. For instance, GIT converts such Japanese names as "信浩", "貴志" and "一幸"to "Nobuhiro", "Takashi" and "Kazuyuki" respectively. In a similar vein, Arabic names "عبدالرحمن" and "عمر" will be converted to "Abdulrahman" and "Omar".

In Step 4, we transform English-alphabet names into phoneme sequences. A phoneme is the smallest set of a language's phonology. For example, John can be represented as the 3-phoneme sequence "JH AA N", while Mary can be represented as "M EH R IY". Our phoneme set contains 40 phonemes that may carry three different lexical stresses, namely no stress, primary stress and secondary stress[19]. We employ the LOGIOS lexicon tool for converting names to phonemes[20]. In this way, we reduce number of features from the order of millions, as in the work of Burger et al. [1], to only around few thousand features, considering all phoneme combinations, and some cases only 40 features. We apply the n-gram analysis to the resulting phonemes. In Section 4 we will compare our phoneme-based method with the word-based (traditional) n-gram feature model used by other authors.

Finally, in Step 5 we analyze our feature sets using KNIME. In general, we observed empirically that the phoneme technique is beneficial to the accuracy of our gender predictions. In general, our accuracy has improved by up to 32.5% from a 50% baseline because of this procedure. We tried both finer and coarser representations for names and we found that phonemes give us the best prediction accuracy among the options that we considered, along with a dramatic reduction in the size of our feature spaces.

IV. Empirical analysis

In this section we evaluate empirically our dataset using different classifiers and we report our findings.

A. Datasets

Upon registering on the Twitter web site, a new Twitter user is presented with a form requesting various kinds of demographic information. However, many of the fields in the form are optional, and indeed a substantial portion of Twitter users leaves many or all of those optional fields blank. In addition, Twitter profiles do not include a specific "gender" field, which complicates our gender identification efforts.

In a Twitter profile, we are interested in the following seven optional fields: (1) First name, (2) user name, (3) profile background color; (4) text color; (5) link color; (6) sidebar fill color; and (7) sidebar border color. Users can either select their own preferences or use the Twitter defaults.

We ran our crawler between February and June 2013. We started our crawler with a set of random profiles and we continuously added any profile that the crawler encountered. Here we follow Alowibdi et al. [3] by filtering all the profiles with valid URLs to reach an explicit gender field.

In all, the dataset consists of 194,293 Twitter profiles, of which 104,535 are classified as male and 89,758 are classified as female. Here we are considering only profiles for which we have obtained gender information independently of Twitter content (i.e., by following links to other profiles). For each profile in the dataset, we collected the seven fields listed above. We also stratified the data by randomly sampling 180,000 profiles, of which about 90,000 are classified as male and about 90,000 are classified as female. In this manner, we obtain an even baseline containing 50% male and female profiles.

Information harvested from Twitter is further processed in various ways. On the one hand, for the colors we use color quantization and sorting as defined by Alowibdi et al. [3]. On the other hand, first names contained in profiles harvested from Twitter undergo a series of pre-processing steps. These steps include the removal of leading and trailing white space, as well as the deletion of last names, numbers, punctuation, and stop words (e.g., Dr, Doc, Mr, Ms). The outcome of this step is first names alone, which can then be used for phoneme sequence generation. The next stage involves computing the phoneme sequences for the preprocessed first names and user names. Phoneme sequences are obtained from LOGIOS and, for profiles in different alphabets, GIT. Next, we generate n-grams of the phoneme sequences. These n-grams and the colors are the feature set input to the classifier. The classifier's empirical results are reported below.

B. Empirical results

We performed different sets of experiments in an effort to assess the relative strengths of the various classifiers. As an additional goal, we wished to assess the effectiveness of our techniques for preprocessing our data set. For this reason, we ran experiments without color quantization and sorting alongside experiments with color quantization and sorting. These results are shown in Table I. In a similar vein, we ran experiments in which we did not transform first names and user names into phoneme sequences. In these cases, we generated n-grams directly from the first names and user names harvested from Twitter. We compared the results obtained in this manner with results obtained by transforming those names into phoneme sequences. These results are shown in Table II and Table III.

We performed different sets of experiments by applying three different classifiers, namely Naïve Bayes (NB), Decision Tree (DT) and Naïve-Bayes Decision-Tree (NB-Tree) hybrid. In all cases, we performed a 10-fold cross validation on data subsets for each classifier. In each set of experiments, we trained our classifiers both with the phoneme-based feature set and with word-frequency based feature set.

We note at the outset that an advantage of the phoneme-based feature set is the reduction in the number of features to a minimum of 40 features, the phoneme set obtained from the LOGIOS lexicon tool, from millions of features in the word-frequency-based method. This reduction results in low computational complexity and a high degree of scalability for the phoneme-based feature set. As we will see, we also obtain reasonably high accuracy results—in the best case 78.5%—even with the small feature set (40 features). An additional advantage of the phoneme-based feature set is language independence, as we obtain phonemes from any language and alphabet system in our dataset. In contrast with our phoneme-based method, the n-gram approach based merely on word frequencies is language dependent while using high dimensional spaces with millions of features generated from unique words extracted from text (i.e. first names and user names).

Table I reports the performance of our dataset using different classifiers for color-based features with a 50.0% baseline. The last five columns in the table report accuracy results for different numbers of color features. Following established practice [8], we define "accuracy" to be the percentage of correctly guessed male users with respect to the total number of male guesses. We use the color features in the order that we listed above, which is the same as the case of our previous report [3]. Thus, the column with one color feature reports only data obtained with the background color alone; the column with two color features reports data for the background color and text color; the next column adds the link color; and the last two columns add sidebar fill and border colors. In fact, we show a better accuracy than our previous results [3], with a 3% improvement resulting from a growth in the size of our dataset from about 55,000 profiles to 180,000 profiles. Our best accuracy for colors alone is 74%, obtained with the NB-Tree classifier.

TABLE I. ACCURACY OF GENDER CLASSIFICATION FOR PROFILE COLORS.

	1 color	**2 colors**	**3 colors**	**4 colors**	**5 colors**
Without Applying Color Quantization and Sorting					
NB	58.0	59.3	61.1	61.1	61.2
DT	**58.9**	**61.2**	63.3	63.1	63.3
NB-Tree	58.0	60.3	**64.7**	**66.2**	**65.7**
With Applying Color Quantization and Sorting					
NB	60.2	61.0	62.0	62.0	63.0
DT	59.0	62.8	65.3	65.0	64.7
NB-Tree	**70.3**	**71.0**	**73.2**	**73.7**	**74.0**

Quantization and sorting of colors result in a significant increase in accuracy, especially when all five-color features are used with the NB-Tree classifiers. In fact, this classifier obtains overall accuracy results of 74% when quantization and sorting are used. Without quantization and sorting this classifier achieves only 65.7% accuracy. Modest performance gains are obtained also with the DT classifier and NB

Figure 1. Cloud tagging of phonemes of male users (left-hand side) and female users (right-hand side).

classifier. The top five colors chosen by female users are Pink, Yellow, Green, Red and Light Blue. The top five colors chosen by male users are Black, Brown, Orange, Gray and Dark Blue.

For first names and user names, we compared our phoneme-based technique with the word frequency method. Without phonemes, we reached around half million features. The size of this feature set is consistent with the results reported by Burger et al. [2]. When using phonemes, the maximum theoretical feature set size for 3-grams is $40^3 = 64{,}000$ features because there are 40 phonemes. However, the largest feature set size that we have observed in practice is around 16,000 phonemes because many phoneme combinations never occur in a 3-gram. Figure 1 shows the cloud tagging of phoneme names for both male and female users. Phonemes in the darker shade of blue are used more frequently than the case of the lighter shade.

When using word frequencies, we conducted experiments with 1-gram through 5-gram features. When using phoneme-based features, we conducted experiments with 1-gram through 3-gram features. TABLE II shows our empirical results for both cases. Entries labeled "NA" refer to cases that were not applicable in our experimental setup. For instance, the name John can be represented as the 3-phoneme sequence "JH AA N" which supports at most a 3-gram analysis. The highest accuracy we obtained was 82.5% in the case of 3-gram phoneme-based features, an improvement of 32.5% with respect to the baseline. In this case, our feature set size was about 16,000 features. The worst-case accuracy for the phoneme-based feature set was predictably the 1-gram case. Even so, we achieved 78.5% accuracy, an improvement of 28.5% over the baseline with only 40 features.

Our accuracy results for phoneme-based gender classification are in line with the methods of Burger et al. [2] and Liu et al. [4]. Those methods obtained an improvement accuracy of 34%, with half a million features, and of 20% with an unknown number of features. Our big advantage is that we obtained accuracy results comparable to their best results with about 16,000 total features. A portion of these features included 10,500 male and female first names available from the US Census Bureau (census.gov/genealogy/names).

Table III indicates that phonemes also work well for user names. The best results were obtained with 3-gram phonemes resulting in 75.2% accuracy and a feature set size of 1,235 features. Evidently, our improvement accuracy over the 50% baseline is 25.2%, which is lower than the accuracy of Burger et al. [2] by about 2%. Thus, the method of Burger et al. [2] is slightly superior to ours with respect to accuracy performance whereas our method is superior to theirs in terms of computational complexity.

Similar to Table II, the data in Table III shows a significant improvement in accuracy for the phoneme-based feature set with respect to the word-frequency based set. The improvement in accuracy is quite significant considering also the lower computational complexity and language independence of the phoneme-based feature set.

TABLE II. ACCURACY OF GENDER PREDICTIONS FOR PROFILES' NAME.

	1-gram	**2-gram**	**3-gram**	**4-gram**	**5-gram**
Without phonemes (n-gram applied to characters of names)					
NB	NA	65.3	67.0	69.2	75.1
DT	NA	68.2	69.3	72.0	76.3
NB-Tree	NA	**69.3**	**70.7**	**74.0**	**78.3**
With phonemes (n-gram applied to set of phonemes)					
NB	65.2	65.3	66.0	NA	NA
DT	**78.5**	**79.2**	**82.5**	NA	NA

TABLE III. ACCURACY OF GENDER PREDICTIONS FOR USER NAMES.

	1-gram	**2-gram**	**3-gram**	**4-gram**	**5-gram**
Without phonemes (n-gram applied to characters of names)					
NB	NA	55.3	56.0	57.2	58.0
DT	NA	**55.7**	**56.9**	**58.2**	**59.6**
NB-Tree	NA	53.2	54.0	56.0	58.0
With phonemes (n-gram applied to set of phonemes)					
NB	55.2	56.0	55.0	NA	NA
DT	**68.5**	**70.2**	**75.2**	NA	NA

On the whole, the accuracy results achieved with first names are higher than the accuracy results obtained with colors and user names. The accuracy of colors and user names are comparable to each other. In the future, we plan to explore accuracy results obtained by combining all three profile

characteristics. In addition, we observe that our phoneme-based n-gram analysis benefits from the addition of features in the 2-gram and especially the 3-gram analysis with respect to the 1-gram analysis. We also note that phonetic analysis of first names and user names can significantly increase our accuracy results. See, for instance, the data relative to the DT classifier in Table II and Table III. Finally, we observe that different classifiers work best with different feature characteristics. In the case of colors, the NB-Tree classifier shows the highest accuracy results. In the case of first and last names, it is the DT classifier that shows the highest accuracy results.

C. Threats to validity

There are two main threats to the validity of this study. The first threat is our reliance on self-declared gender information entered by Twitter users on external web sites for validation of our predictions. We use this gender information as our ground truth. Evidently, a complete evaluation of all 194,293 Twitter users would be impractical. We manually "spot-checked" about 5,000 out of the 194,293 profiles in our dataset or about 2.5% of the dataset. In the cases that we checked by hand, we are confident that the gender information we harvested automatically was indeed correct. The second threat is given by the overall size of the dataset that we could analyze. Although we started from four millions Twitter users, we ended up with just 194,293 users whose gender we could verify independently. This indicates that the size of the training sets was adequate; however, we will continue expanding our data set.

V. Conclusion and future work

In this paper, we empirically studied gender classification on Twitter using different profile characteristics such as first name, user name, background color, text color, link color, sidebar fill color, sidebar border color. Also, we presented a novel approach to predict gender utilizing phoneme-based features extracted from profile names and user names. In addition, we applied both finer and coarser representations for first names, user names and colors.

The main advantage of our gender-classification methods is that they achieve good accuracy results, despite sharp reductions in computational complexity with respect to alternative approaches. Our methods also have broad applicability to different languages and alphabet sets than English. In the future, we intend to apply additional profile and tweet characteristics, such as the content of tweets, for gender classification. We also plan to investigate combinations of characteristics in order to improve our prediction accuracy even further.

References

[1] R. Wauters, "Only 50% Of Twitter messages are In English, study says", TechCrunch.com, http://techcrunch.com/2010/02/24/twitter-languages/, February 2010, accessed in October, 2013.

[2] L. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating gender on Twitter". In *Proceedings of* EMNLP'11 1301–1309, 2011.

[3] J. Alowibdi, U. Buy and P. Yu, "Language Independent Gender Classification on Twitter", *The 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM'13*, Niagara Falls, Canada, 2013.

[4] W. Liu, and D. Ruths, "What's in a name? Using first names as features for gender inference in Twitter", In *Symposium on Analyzing Microtext*, 2013.

[5] F. Al Zamal, W. Liu, D. Ruths, "Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors". *Int'l AAAI Conference on Weblogs and Social Media*, 2012.

[6] W. Liu, F Al Zamal, D. Ruths, "Using Social Media to Infer Gender Composition of Commuter Populations". *Int'l AAAI Conference on Weblogs and Social Media,* 2012.

[7] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in Twitter". In *Proceedings of SMUC'10*, 37–44, 2010.

[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA Data Mining Software: An Update", *SIGKDD Explorations*, Vol. 11, Issue 1, 10–18, 2009.

[9] M. Berthold, N. Cebron, F. Dill, T. Gabriel, T. Kotter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel, "KNIME—The Konstanz information miner: Version 2.0 and beyond" *SIGKDD Explor.* 26-31, 2009.

[10] S. Singh, "A pilot study on gender differences in conversational speech on lexical richness measures". In *Literary and Linguistic Computing Journal*, vol. 16, 251–264, 2001.

[11] S. Argamon, M. Koppel, J. Fine, and A.R. Shimoni, "Gender, Genre, and Writing Style in Formal Written Texts". *Text*, vol. 23, no. 3, 321–346, 2003.

[12] S. Herring, L. Scheidt, S. Bonus, and E. Wright, "Gender and genre variation in weblogs". *Journal of Sociolinguistics*, 439–459, 2006.

[13] T. Kucukyilmaz, B.B. Cambazoglu, C. Aykanat, and F. Can, "Chat mining for gender prediction". In *Proceedings of the 4th ADVIS'06*, 274–283, 2006.

[14] C. Peersman, W. Daelemans, and L. Vaerenbergh, "Predicting age and gender in online social networks". In *Proceedings of SMUC'11*, 37–44, 2011.

[15] R. Sarawgi, K. Gajulapalli, and Y. Choi, "Gender attribution: Tracing stylometric evidence beyond topic and genre". In *Proceedings of CoNLL'11*, 78–86, 2011.

[16] M. Koppel, S. Argamon, and A. Shimoni, "Automatically Categorizing Written Texts by Author Gender". *Lit Linguist Computing*, 401–412, 2002.

[17] A. Mukherjee, and B. Liu, " Improving gender classification of blog authors". In *Proc. EMNLP'10*. 207–217, 2010.

[18] S. Nowson, J. Oberlander, and A. Gill, " Gender, Genres, and Individual Differences". *In Proc. of the 27th annual meeting of the Cognitive Science Society,* 1666–1671, 2005.

[19] Speech at CMU, "The CMU pronouncing dictionary", Speech lab, http://www.speech.cs.cmu.edu/cgi-bin/cmudict/, Oct. 2007, Accessed in Oct. 2013.

[20] Speech at CMU, "LOGIOS Lexicon tool", Speech lab, http://www.speech.cs.cmu.edu/tools/lextool.html, Oct. 2007, Accessed in Oct. 2013.