Language Independent 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) generate a huge volume of user-originated texts. Gender classification can serve multiple purposes. For example, commercial organizations can use gender classification for advertising. Law enforcement may use gender classification as part of legal investigations. Others may use gender information for social reasons. Here we explore language independent gender classification. Our approach predicts gender using five color-based features extracted from Twitter profiles (e.g., the background color in a user's profile page). Most other methods for gender prediction are typically language dependent. Those methods use high-dimensional spaces consisting of unique words extracted from such text fields as postings, user names, and profile descriptions. Our approach is independent of the user's language, efficient, and scalable, while attaining a good level of accuracy. We prove the validity of our approach by examining different classifiers over a large dataset of Twitter profiles.

***Keywords*- Color-based Features, Social Network, Application for Social Network, Language Independent.**

I. Introduction

Online Social Networks (OSNs) play a significant role in the daily life of many people and organizations. The onset of OSNs has stretched the traditional notion of "community" to include groups of people who never met in person but communicate with each other through OSNs to share knowledge, opinions, interests, activities, relationships, and friendships. The key factor underlying the success of OSN-mediated communities, similar to traditional communities, is the trust that exists among community members.

Gender classifications typically are language dependent, not scalable, inefficient, held offline using high-dimensional spaces. A recent study [1] shows that there are around 78 different languages in Twitter with English as the dominant language. Another study by Wauters shows that only around 50% of Twitter messages are in English[1]. Our Twitter dataset alone contains 31 different languages. An estimate breakdown of language use in our dataset shows that around 82% users are English with the remaining 18% distributed over 30 languages. Most existing research for gender classification on Twitter is language dependent. A recent study for gender classification [2] shows that 66% of users in their dataset use English. Other works for gender classification [3], [4], [5], did not mention the language distribution of their Twitter dataset, which we assume to be in English. On the whole, our work is different from existing methods in term of its simplicity, language independence and low computational space and time complexity.

Here, we report our results on language-independent gender identification based on profile colors alone. In particular, we predict automatically the gender value of users based on their color 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 [15], [16]. Unlike text-based approaches, we present a novel method for predicting gender using five color-based features. Our main contributions are outlined below.

- We defined a novel approach for predicting gender using color-based features.
- Our method is language independent; most other existing methods that use text are restricted to one language or few languages.
- We validated our approach by analyzing different classifiers over a large dataset of Twitter profiles. Our results show that colors alone can provide reasonably accurate gender predictions.
- We defined a color quantization and sorting technique for preprocessing colors harvested from Twitter profiles. This technique substantially improves prediction accuracy.
- Colors alone are not useful features. However, we found that considering a combination of multiple (five) color selections from each Twitter profile leads to a reasonable degree of accuracy for gender prediction.
- Unlike existing methods that use millions of features and high-dimensional spaces, we only utilize five color-based features and a low-dimensional space. As a result, our color-based analysis is quite promising in terms of computational complexity compared to other gender-guessing methods, which use much larger feature sets.

The remainder of this paper is organized as follows. In Section 2, we detail our proposed approach. In Section 3, we report our empirical results from different classifiers and we analyze these results. In Section 4, we briefly summarize related work on gender classification. Finally, in Section 5, we draw final conclusions and outline future work.

[1] Wauters, R. http://techcrunch.com/2010/02/24/twitter-languages/, access in May 2013.

II. Proposed approach

We harvest colors from user profiles. Next, we apply a color reduction and quantization procedure (i.e., normalization) to reduce the number of colors. The colors are converted from their Red, Green and Blue (RGB) representation to the corresponding HSV (Hue, Saturation, Value) representation. We then sort the colors by their hue and value, and finally we convert them back to RGB. The sorting allows labeling similar colors (e.g., adjacent colors in the sort) by consecutive numbers that we feed to the classifier.

Colors harvested from Twitter user profiles are typically specified as a combination of RGB values ranging between 0 and 255. This gives a total of 256^3 colors combinations. Because of the large number of combinations, we use quantization, a compression procedure that substantially reduces the huge number of colors. Each of the red, green and blue values is shrunk from 8 bits to 3 bits. This technique reduces the total number of color combinations from $256^3 \approx 16 \cdot 10^6$ to just $8^3 = 512$ colors. Each of the original colors we harvested is converted to the compressed color having the least Euclidean distance from the original color. Next, we convert each quantized color to the corresponding HSV representation. We use this representation for sorting the colors according to their similarity. First, colors are sorted by their hue; we use values to break ties between colors having identical hues.

We observed empirically that quantization and sorting are beneficial to the accuracy of our gender predictions. In general, our accuracy has improved by up to 13% because of these procedures. We tried both finer and coarser representations for colors and we found that 3 bits per color give us the best prediction accuracy among the options that we considered. We conclude that this representation is a reasonable compromise between the number of colors (i.e., the feature values) that we must consider and the perceptual differences within the resulting color clusters. Color quantization is especially important because we are using a total of 5 color features for each user we analyze. In general, quantization reduces the number of cases (i.e. combinations) for five color-based features from 256^{3*5} cases to 8^{3*5} cases.

III. Experimental results

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

Datasets

We chose Twitter profiles as the starting point of our data collection. A Twitter user must first fill a profile form, consisting of about 30 fields containing biographical and other information, such personal interests and hobbies. However, many fields in the form are optional, and indeed substantial portions of Twitter users leave many or all of those optional fields blank. In addition, the Twitter's profile form does not include a specific "gender" field, which complicates gender identification for Twitter users.

Like many other fields in a Twitter profile, here we are interested in the five fields that allow users to choose different colors for the following items: (1) Background color; (2) Text color; (3) Link color; (4) Sidebar fill color; and (5) Sidebar border color. Users choose their own preferences by selecting colors from a color wheel while editing their profiles. Unlike other OSNs, such as Facebook, Twitter allows users to redesign and change their profiles.

We ran our crawler between March and May 2012. We started our crawler with a set of random profiles and we continuously added any profile that the crawler system encountered (e.g., profiles of users whose names were mentioned in tweets we harvested). Subsequently, we filtered all the profiles with valid URLs. The URL is a profile field that lets a Twitter user create a link to a profile hosted by another OSN, such as Facebook. This field is important because profiles hosted by other OSNs often contain an explicit gender field.

In all, the dataset consists of 53,326 profiles, of which 30,898 are classified as male and 22,428 are classified as female. We are considering only profiles for which we have obtained gender information independently of Twitter content. For each profile in the dataset, we collected the five profile colors listed above. We are considering all these colors in our empirical study on gender classification.

Twitter offers 19 predefined designs, including a default design, to each new user joining the social network. Each design defines colors for all five fields. Users can select those designs easily. For instance, Twitter offers the color (R=192, G=222, B=237), a light shade of blue, as the default background color to any new user.

In order to account for the existence of predefined designs in the Twitter user setup, we have considered different subsets of our overall dataset, and we studied each subset independently of other subsets. We specifically considered the following subsets:

T1. This is the entire dataset, *A*, consisting of 53,326 profiles with a 57% male and 43% female breakdown.

T2. This is dataset *A-D,* which is the subset containing all collected profiles, except for profiles using the default design with the RGB values of (192, 222, 237) as the background color, denoted by *D*. *D* represents 16% of dataset *A* while T2 represents 84%. The base condition is male with a 55% representation.

T3. This is dataset is *A-C,* which is the subset obtained by excluding *C*, the subset all profiles that use any of the 19 predefined designs including the default design, from *A*. *C* represents 57% of *A* while T3 represents 43%. The base condition is female with a 50.5% percentage. Here we report detailed empirical results about T3, since it includes only profiles with custom color choices, and we summarize results for the other datasets.

T4. This is dataset *A-B,* obtained by excluding from the entire dataset, *A*, all profiles, *B*, that use any of the 19 predefined designs as well as black or white as background color. *B* represents 71% of *A*, while T4 represents 29%. The base condition is female with a 54.5% percentage.

Figure 1 shows the five subsets that we considered for our analyses. Overall, female users are more likely to choose

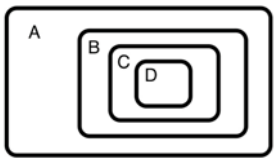

Figure 1 Four Subset of our dataset.

their own layout colors, while male users are more likely to use the default design or one of the other predefined designs.

Empirical studies

We performed four sets of experiments, one for each of the five subsets of our dataset. In each experiment set, we applied four different classifiers, namely Probabilistic Neural Network (PNN), Decision Tree (DT), Naïve Bayes (NB) and Naïve Bayes/Decision-Tree Hybrid (NB-Tree). We performed a 10-fold cross validation on our five data subsets for each classifier. In each set of experiments, we trained our classifiers with all five color-based features.

We assessed the effectiveness of color quantization by running experiments with and without color quantization (i.e., using the raw RGB data harvested from the Twitter profiles). TABLE I and TABLE II report the performance of dataset T3 using different classifiers and color-based features with a 50.5% female baseline. The last five columns in the table report results for different numbers of color features. We use the color features in the order that we listed previously. 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. For each experiment, we report the percentage of correctly identified male users and female users and the overall accuracy.

On the one hand, TABLE I reports the accuracy of gender prediction. The quantization and sorting algorithms discussed above are not applied in this case. On the other hand, the data in TABLE II was obtained after applying quantization to Twitter profile colors and sorting the resulting color clusters. As shown in TABLE I without quantization, the performance of three color-based features roughly equals the case of four and five features. In the case of the PNN classifier, three features actually give better accuracy than four and five features. In contrast with TABLE I, in TABLE II, the accuracy performance increases when using all five color-based features compared to the case of three color-based features.

On the whole, the data in TABLE I and TABLE II show that quantization and sorting of colors result in a significant increase in accuracy, especially when all five-color features are used with the Probabilistic Neural Network (PNN) and Naïve Bayes/Decision-Tree Hybrid (NB-Tree) classifiers. In fact, these two classifiers obtain overall accuracy results of 69% and 71.4% with quantization and sorting. Without quantization and sorting these two classifiers achieve only 59.4% and 65.7% accuracy. Modest performance gains are obtained also with the Decision Tree (DT) classifier. In contrast with the other three classifiers, the Naïve Bayes (NB) classifier fails to achieve any gains. In fact, the performance of this classifier drops with color quantization and sorting.

Figure 2 shows the accuracy increase obtained by using the color quantization procedure compared to the case of raw RGB colors for each of the four classifiers on dataset T3. Part (a) shows the performance of the Naïve Bayes classifier with and without quantization. This is the only classifier that provides slightly better accuracy without quantization than in the case of quantization. However, the overall performance of the classifier is inferior to that of the other classifiers. Part (b) in Figure 2 shows the performance of the Decision Tree classifier, which yields better accuracy than Naïve Bayes. In this case, color quantization and sorting improve slightly the accuracy of the predictions. The performance of the Probabilistic Neural Network (PNN) and Naïve Bayes/Decision-Tree Hybrid (NB-Tree) classifiers are shown in Part (c) and Part (d) of Figure 2.

TABLE III shows the performance of the four classifiers on all four datasets that we considered after color quantization. Evidently, NB-Tree has the best accuracy on all five datasets with accuracy results consistently above 70% in all four cases. We specifically obtained our best results with the NB-tree classifier in the T3 dataset with an accuracy of 71.4% over a 50.5% female baseline, a gain of about 20%.

An advantage of our approach is that uses only five colors, making it language independent. An additional advantage is that it has a low-dimensional space, resulting in a low computational complexity of our classifiers. In contrast with our method, most existing approaches are language dependent while using high dimensional spaces generated from unique words extracted from text (i.e. tweets, names, and profile descriptions), and millions of features. For instance, Burger et al. [2] utilize 15.6 million features with each feature corresponding to a unique word extracted from a tweet. Similarly, Rao et al. [5] use 1.25 million features extracted from tweets.

Figure 3 shows the effects of different training set sizes on the accuracy of the predictions. Similar to Figure 2, the four parts of the figure refer to different classifiers; for each classifier we use color-coded lines to distinguish the number of color features that we consider.

TABLE I. ACCURACY OF GENDER PREDICTIONS FOR DATASET T3 WITH RGB COLORS WITHOUT QUANTIZATION.

	1 color	2 colors	3 colors	4 colors	5 colors
NB	58.0	59.3	61.1	61.1	61.2
DT	58.9	61.2	63.3	63.1	63.3
PNN	**60.0**	**64.2**	**67.0**	63.3	59.4
NB-Tree	58.0	60.3	64.7	**66.2**	**65.7**

TABLE II. ACCURACY OF THE EXPERIMENT RESULTS FOR DATASET T3 AFTER APPLYING COLOR QUANTIZATION AND SORTING.

	1 color	2 colors	3 colors	4 colors	5 colors
NB	49.9	56.8	58.2	56.6	56.0
DT	**59.0**	62.8	65.3	65.0	64.7
PNN	**59.0**	63.4	67.0	68.1	69.0
NB-Tree	58.3	**64.4**	**70.2**	**71.3**	**71.4**

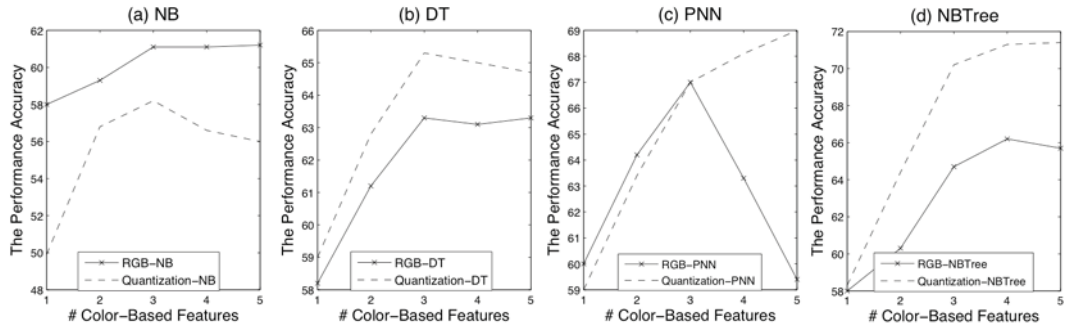

Figure 2. Accuracy of the four classifiers on dataset T3 using different numbers of color-based features .

All diagrams refer to dataset T3. In general, the larger training sets yield better accuracy results; however, the Naïve Bayes classifier again differs from the other three classifiers in that its best results are obtained with smaller training sets. The other three classifiers show similar behaviors with respect to training set size. When one or two color features are considered, the performance grows steadily with the size of the training set. However, when three or more color features are considered, the growth in performance levels off between 5,000 and 10,000 profiles, or it is essentially flat (in the case of the NB-Tree classifier). These results show that the size of our training sets is adequate.

Figure 4 shows the difference in colors chosen by female vs. male Twitter users. On the left-hand side we show popular colors chosen by female users (after clustering); the colors for male users are shown on the right-hand size of the figure.

TABLE III. ACCURACY OF THE EXPERIMENTAL RESULTS FOR THE FIVE DIFFERENT DATASETS WITH COLOR QUANTIZATION AND SORTING.

	Scores (%)	T1	T2	T3	T4
NB	Precision	67.4	63.5	67.0	63.1
	Recall	71.0	68.1	60.2	58.2
	F-score	69.2	65.7	63.4	60.6
	Accuracy	65.2	63.5	61.7	62.6
DT	Precision	81.6	77.0	66.0	61.4
	Recall	70.2	69.0	63.9	60.0
	F-score	75.5	72.7	65.0	60.7
	Accuracy	69.3	68.2	64.7	63.8
PNN	Precision	**84.0**	**80.5**	71.3	64.1
	Recall	71.8	71.0	**67.8**	**65.5**
	F-score	**77.4**	**75.4**	69.5	64.8
	Accuracy	**71.6**	**71.1**	69.0	68.3
NB-Tree	Precision	76.9	78.2	**83.6**	**80.6**
	Recall	**73.6**	**71.8**	67.0	64.8
	F-score	75.2	74.8	**74.3**	**71.8**
	Accuracy	70.7	**71.1**	**71.4**	**71.2**

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 53,000 Twitter users would be impractical. We manually "spot-checked" about 1,000 out of the 53,000 profiles in our dataset or about 2% of the dataset. In all cases that we checked by hand, we are confident that the gender information we harvested was indeed correct. Thus, we are confident that the gender information for the entire dataset is quite accurate. The second threat is given by the overall size of the dataset that we could analyze. Although we started from one million Twitter users, we ended up with just 53,000 users whose gender we could verify independently. However, the data in Figure 3 indicates that the size of training sets was at least adequate. Apparently, little will be gained by using larger datasets.

IV. Related work

Many researchers have investigated gender classification. Lexical richness measures based on word-frequencies have also been studied [6]. Also, Argamon et al. [7] defined a POS n-gram technique to capture author's writing styles. POS tags, unigrams, word-frequencies, word-classes, POS patterns, POS contents and POS style metrics have been studied by many authors [8], [9], [10], [11], [12], [13], [14]. Unlike those works, Burger et al. [2] and Rao et al. [5] worked on gender classification on Twitter postings by utilizing text sentiment. In particular, Rao et al. [5] use sociolinguistic-feature models, Ngram-feature models and stacked models for gender classification utilizing text sentiment. Burger et al. use the Ngram-feature model [2]. There are millions of features generated using text sentiments in both approaches.

In summary, most existing authors explore gender classification by utilizing the text sentiment approach. Researchers in the natural language processing and data mining communities worked on gender classification of different systems including OSNs for the past several years. Despite the challenging feature set of those systems, researchers have studied various schemes for defining feature feasibility and stability. In general, researchers must pay more attention to the structure of those systems. Although most existing research ex-

Figure 4. Spectrum of popular colors for female users (left-hand side) and male users (right-hand side).

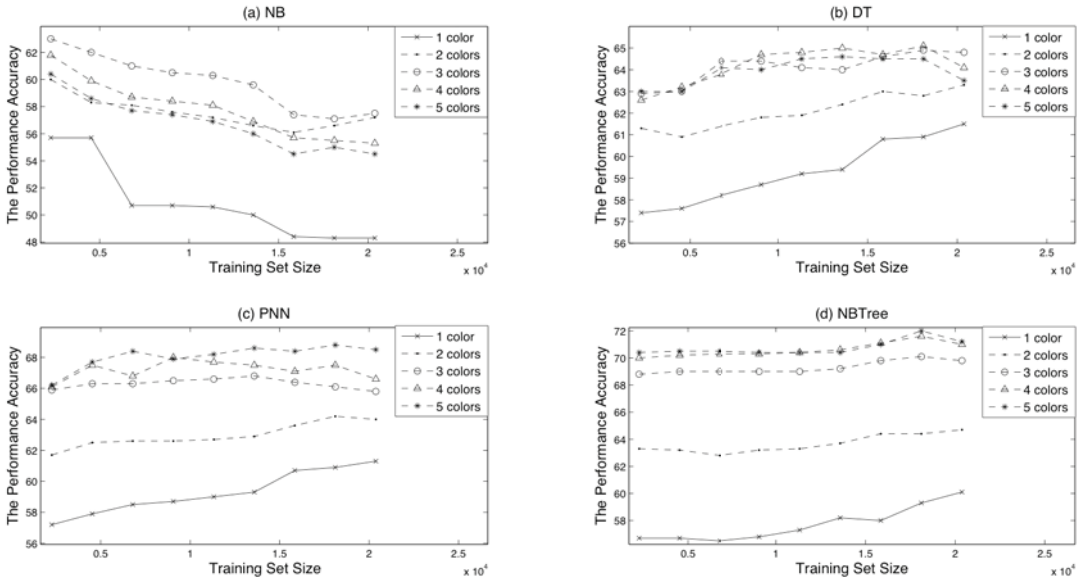

Figure 3. Effects of different training set sizes on accuracy of different classifiers on dataset T3 with different numbers of color-based features.

tracted millions of features from text sentiment based on the structure of the systems, our work shows that reasonably accurate predictions are possible using only five color-based features. The drawback of using text sentiment is high computational complexity for the generated high dimensional space, language dependency and millions of features.

V. Conclusion and future work

In this paper, we studied gender classification on Twitter. We presented a novel approach to predict gender utilizing only five color-based features extracted from the profile layout colors (i.e. background). Our approach is independent of the user's language, held online, scalable, efficient and has low computational complexity, while attaining a reasonable level of accuracy. We prove the validity of our approach by examining different classifiers over a large dataset of Twitter profiles.

In the future, we intend to study different characteristics of the dataset to classify gender (e.g., features of a user's friends and followers, names, screen names, description) and to incorporate them in our framework to detect deception.

References

[1] D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, et al., "The Twitter of Babel: Mapping World Languages through Microblogging Platforms", 2013.

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

[3] 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.

[4] 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,* 2012.

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

[6] 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.

[7] 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.

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

[9] 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.

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

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

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

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

[14] 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.

[15] 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.

[16] 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.