Say It with Colors: Language-Independent Gender Classification on Twitter

Jalal S. Alowibdi [1,2], Ugo A. Buy [1], and Philip S. Yu [1]

[1] Department of Computer Science, University of Illinois at Chicago
Illinois, USA
[2] Faculty of Computing and Information Technology, King Abdulaziz University
Jeddah, Saudi Arabia
jalowibd, buy, psyu@cs.uic.edu

Abstract. Online Social Networks (OSNs) have spread at stunning speed over the past decade. They are now a part of the lives of dozens of millions of people. 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. Here we explore in depth language independent gender classification. Our approach predicts gender using five color-based features extracted from Twitter profiles such as the background color in a user's profile page. This is in contrast with most existing methods for gender prediction that are 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, scalable, and computationally tractable, while attaining a good level of accuracy.

Keywords: Color-based feature, gender classification, Twitter profile, color quantization, color feature, low dimensional space, social network analysis, online social network.

1 Introduction

Online Social Networks (OSNs) generate a huge volume of user-originated texts. OSNs allow users to share knowledge, opinions, interests, activities, relationships and friendships with each other. Gender classification can serve multiple purposes in these settings. 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 examine gender classification based solely on color preferences. We specifically present a novel approach for predicting gender using five color-based features extracted from Twitter profile colors (e.g., the background color in a user's profile page) that is.

Methods for gender classification are typically language dependent, not scalable, inefficient, and 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 [2] shows that only around 50% of Twitter messages are in English. Our Twitter dataset alone contains 34 different languages. An estimate breakdown of language use in our dataset shows that around 69% users are English speaking with the remaining 31% distributed over 33 languages. In addition, around 20% of the 69% users who set their profiles to be as English speaking, routinely post texts in different languages than English. Thus, about 45% of users in our dataset use languages different than English for their posts and profiles. Our long-term goal is gender identification in OSNs with an emphasis on accuracy, computational efficiency and scalability of gender predictions. We are especially interested in language-independent methods.

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 profile colors for gender classification is currently unknown. Most existing research for gender classification on Twitter is language dependent. An existing study for gender classification [3] shows that 66% of users in their dataset use English. Other works for gender classification [4], [5], [6] did not mention the language distribution of their Twitter dataset, which we assume to be in English. In contrast, our dataset contains profiles of users of all ages, languages, and cultures. In particular, Burger et al. [3] used four different characteristics from a user's profile and posts (i.e., first name, user name, description and tweets) for gender classification. Liu and Ruths [7] utilized only first names for gender classification. Alowibdi et al. [8] applied a phoneme-based analysis to characteristics extracted from a user's profile (e.g., first names and user names). Other works for gender classification use user posts and other statistical information, such as friends and followers, in order to identify gender [4], [5], [6], [9]. In general, 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 [3]. On the whole, our work to predict gender from profile's colors is unique and different from existing methods in term of its simplicity, language independence and low computational space and time complexity. In addition, our work is different because of the range of profile colors characteristics that we consider.

We predicted 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 [10], [11]. Unlike text-based approaches, we used a novel method for predicting gender using five color-based features. Our preliminary results with our data set are quite encouraging. Although we are considering only five color-based features, we can predict gender with an accuracy of 74.2%, a gain of about 24% with respect to a 50% baseline. A key to the success of our gender guessing with colors is our preprocessing of color features using a quantization technique that we discuss later on. An ad-

vantage of our method is its broad applicability to Twitter users regardless of their language; we use only color-based features to identify gender. In addition, our color-based analysis shows promising results in term of computational complexity compared to other gender-guessing methods, which use a much larger feature set. Our approach utilizes only five color-based features while Burger et al. [3] and Rao et al. [6] use text sentiment with 1.2 million and 15.4 million features. Our results show that colors alone can provide reasonably accurate gender predictions, even though a substantial number of users we analyzed do not change the default colors provided by Twitter in their Twitter profiles or in other web sites hosting their profiles (e.g., Twitter App). We conclude that colors are a good gender indicator for users who do change the default colors in their profiles. In these cases, we will be able to use colors alone as part of our gender classification methods.

Our main contributions are outlined below.

1. We defined a novel, language-independent approach for predicting gender using color-based features. Most other existing methods rely on text, which varies by language.
2. 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. In some cases, we can predict gender with compatible accuracy of 74.2%, a gain of about 24% with respect to a 50% baseline.
3. We defined a color quantization and sorting technique for preprocessing colors harvested from Twitter profiles. This technique substantially improves prediction accuracy while also reducing dramatically the size of our feature set. As a result, our color-based analysis has much lower computational complexity than most other other gender-guessing methods, which use much larger feature sets based on text features.
4. We concluded that 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.

The remainder of this paper is organized as follows. In Section 2, we briefly summarize related work on gender classification. In Section 3, we described our dataset collection. In Section 4, we detail our proposed approach. In Section 5, we report our empirical results from different classifiers and we analyze these results Finally, in Section 6, we give some conclusions and outline future work.

2 Related work

Many researchers have investigated gender classification. Lexical richness measures based on word-frequencies have also been studied [12]. For instance, Argamon et al. [13] defined a POS n-gram technique to capture author writing styles. Many authors have studied POS tags, unigrams, word-frequencies, word-classes, POS patterns, POS contents and POS style metrics [14], [15], [16], [17], [18], [19],

[20]. Unlike those works, Burger et al. [3] and Rao et al. [6] worked on gender classification on Twitter postings by utilizing text sentiment. In particular, Rao et al. [6] use sociolinguistic-feature models, n-gram feature models and stacked models for gender classification utilizing text sentiment. Burger et al. use the n-gram feature model [3]. Both approaches generate millions of features from text sentiments.

In summary, most existing authors explore gender classification by utilizing language-based methods. 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. The drawback of using text sentiment is high computational complexity of the high dimensional space generated, language dependency, and millions of features. Our work shows that reasonably accurate predictions are possible using only five color-based features.

3 Dataset Collection

We chose Twitter profiles as the starting point of our data collection for several reasons. First, Twitter is one of the most popular social networks to date with a huge user community cutting across great many languages, cultures and age groups. In early 2013, Twitter reached 555 million registered users [21]. As of today, Twitter states that there are more than 200 million active users producing around 400 million tweets per a day [22]. Second, Twitter has all the color attributes that we need to set up the experiment. These attributes are generally public, meaning that they can be accessed and viewed by anyone who requests them. Lastly, Twitter provides a rich Application Programming Interface (API), which supports automatic collection of large data sets.

For our experiments, we chose Twitter profiles as the starting point of our data collection. In Twitters terminology, the followers of a given user U are users interested in reading $U's$ tweets. These users will be notified when U posts a new tweet. Also, the friends of a user V are the users following $V's$ tweets. In general, users can register themselves as followers of any other user; no permission is required unless the user protects his/her profile using Twitter's protection features. A new Twitter user must first fill a profile form, consisting of about 30 fields containing biographical and other personal 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, Twitter's profile form does not include a specific "gender" field, which complicates gender identification for Twitter users. One can choose additional fields that are not mentioned above for gender classification such as posted tweets; however, we decided to perform gender classification using only profile colors.

Among 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.
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. In some cases, users chose both a background color and a background picture (from a picture file) for their profiles. In these cases, the background picture overrides the background color, which is not shown. However, our empirical setup will take into account the background color chosen by a user even if that color is overriden by that user.

We ran our crawler between August and December 2013, subject to Twitter's limitation of less than 150 requests per hour. We started our crawler with a set of random profiles and we continuously added any profile that the crawler 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, which Twitter profiles do not include.

In all, the dataset we used at the time of our study consisted of 169,449 profiles, of which 94,251 were classified as male and 75,198 were classified as female. We considered only profiles for which we obtained gender information independently of Twitter content (i.e., by following links to other profiles). For each profile in the dataset, we collected the five profile colors listed above. We also stratified the data by randomly sampling 150,000 profiles, of which about 75,000 are classified as male and about 75,000 are classified as female. In this manner, we obtain an even baseline containing 50% male and female profiles. 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. As of this writing, the color (R=192, G=222, B=237), a light shade of blue, is the default background color for any new Twitter 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. In addition, we stratified each subset by randomly sampling the profiles, from which we obtain even baselines containing 50% male and female profiles. We specifically considered the following subsets:

T1. This is the entire dataset, A, consisting of 150,000 profiles with a 50% male and 50% 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 11.4% of

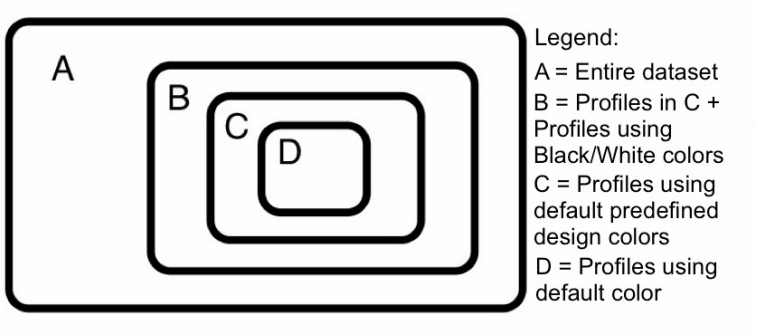

Fig. 1. Four Subset of our dataset

dataset *A* while *T2* represents 88.6%. The base condition is a 50% male and 50% female breakdown.

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 around 57% of *A* while *T3* represents 43%. The base condition is a 50% male and 50% female breakdown. 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 71.8% of *A*, while *T4* represents 28.2%. The base condition is still a 50% male and 50% female breakdown.

Figure 1 shows the four subsets that we considered for our analyses. Overall, female users are more likely to choose their own layout colors, while male users are more likely to use the default design or one of the other predefined designs.

4 Proposed Approach

Our algorithm for preprocessing colors before feeding the colors to the classifier is shown in Figure 2 below. First, we harvest colors from user profiles. Next, we apply a color quantization and sorting 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.

Figure 3 shows the color distribution of profile background colors harvested from profiles in our data set before quantization. Broader stripes denote the relative frequency of background color in the profiles that we analyzed. In particular, the broad light blue stripe to the center left of the figure represents the default background color of Twitter profiles.

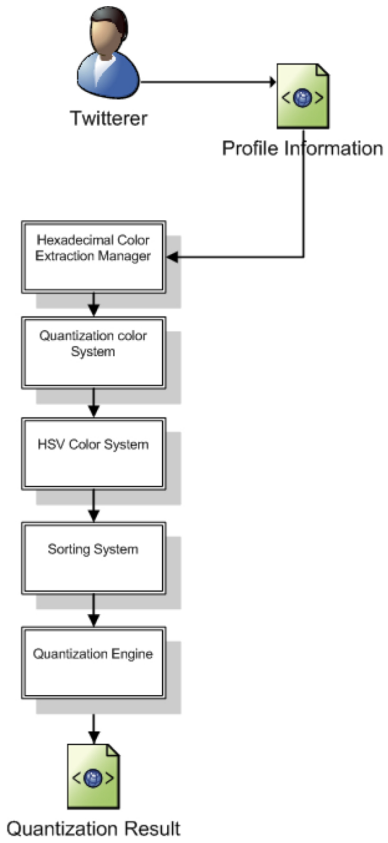

Fig. 2. Algorithm for color preprocessing

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 4 bits and 3 bits respectively. This technique reduces the total number of color combinations from $256^3 \approx 16 * 10^6$ to just $16^3 = 4096$ colors and $8^3 = 512$ colors, respectively. Each of the original colors we harvested is converted to the compressed color having the least Euclidean distance from the original color. Next, according to the algorithm in Figure 2, 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. Figure 4 below shows the 512 colors (i.e., the quantization color procedure of 9-bit RGB) obtained after quantization and sorting.

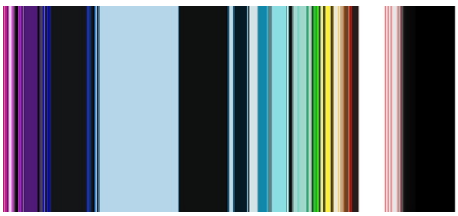

Fig. 3. Distribution of profile background colors before applying color quantization in our dataset

The rationale for applying color quantization is that the feature set obtained from straight RGB values would be quite large, a total of $256^{(3*5)}$ cases for 5 color features. A feature set of this size would be mostly unnecessary as most colors are perceptually indistinguishable from neighboring colors with R, G, and B values differing only by few units from the original color. Thus, we chose to cluster colors in such a way that colors with a given cluster are perceptually similar to each other. Next, we investigated the size of each cluster. Larger clusters would lead to smaller features sets; however, larger clusters may also lead to the inclusion of substantially different colors in the same cluster. For this reason, we studied empirically clusters of various sizes and we concluded that clusters grouping 512 colors in each cluster, with 5-bit RGB values per cluster, gave us the highest accuracy results.

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 15% because of these procedures. Figure 5 shows in 3 dimensions the profile background colors distribution for male and female users, the quantization color centroid and background color distribution for both genders in our data set after applying the quantization color procedure of 9-bit RGB. In brief, our quantization color procedure is a reduction from 24-bit to both 12-bit and 9-bit RGB color representations. 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

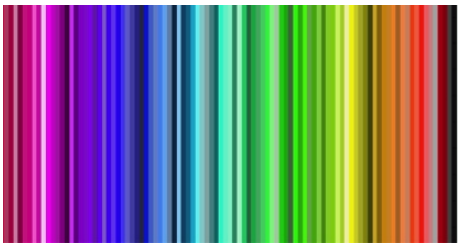

Fig. 4. Spectrum of sorted, quantized colors obtained by the color-preprocessing algorithm shown in Figure 2

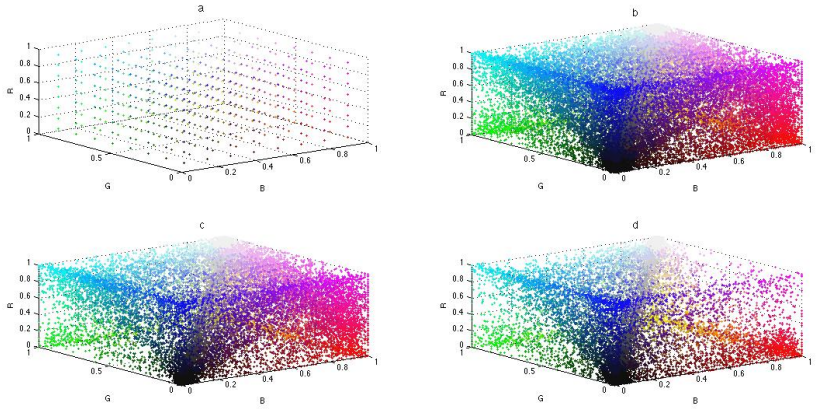

Fig. 5. Part (a) shows the centroid of the quantization color procedure; Part (b) shows the color distribution of both genders for the profile background after applying the quantization color procedure to our data set; Part (c) shows the color distribution of the profile background of female users after applying the quantization color procedure to our data set; and Part (d) shows a similar color distribution for male users

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 $32^{(3*5)}$ cases.

5 Empirical studies

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

5.1 Experimental results

We performed four sets of experiments, one for each of the four subsets of our dataset. In each experiment set, we tried many classifiers; different classifiers produced different results. Next, we selected the top classifiers. Here we consider the following four different classifiers: 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 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

	Scores (%)	1 color	2 colors	3 colors	4 colors	5 colors
NB	Precision	59.2	59.1	61.1	62.1	62.2
	Recall	59.2	59.1	61.1	62.1	62.2
	F-score	59.2	59.1	61.1	62.1	62.2
	Accuracy	59.2	59.1	61.1	62.1	62.2
DT	Precision	59.9	61.5	63.7	64.0	64.1
	Recall	58.8	61.5	63.8	64.0	64.1
	F-score	57.9	61.4	63.8	64.0	64.1
	Accuracy	58.8	61.5	63.8	64.0	64.1
PNN	Precision	**62.2**	**65.6**	**66.7**	66.2	**66.9**
	Recall	**61.2**	**65.7**	**66.5**	65.4	65.0
	F-score	**60.5**	**65.7**	**66.4**	63.2	63.9
	Accuracy	**61.3**	**65.7**	**66.6**	64.4	65.0
NB-Tree	Precision	58.6	61.1	64.4	**67.2**	65.2
	Recall	58.3	61.1	64.4	**67.2**	**65.2**
	F-score	57.9	61.1	64.4	**67.1**	**65.2**
	Accuracy	58.2	61.1	64.4	**67.2**	**65.2**

Table 1. Accuracy of gender predictions for dataset T3 with RGB colors without quantization

from the Twitter profiles). Table 1 and Table 2 report the performance of dataset T3 using different classifiers and color-based features with a 50% baseline. In particular, we choose T3 among the other datasets because T3 is our largest data subset containing only colors chosen by users from the color wheel. 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 1 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 2 was obtained after applying quantization to Twitter profile colors and sorting the resulting color clusters. As shown in Table 1 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. Also, in the case of the NB-Tree classifier, four features actually give better accuracy than three and five features. In the case of the NB-Tree classifier, four features provide the best accuracy for the RGB Colors. In contrast with Table 1, in Table 2 the accuracy performance increases when using all five color-based features compared to the cases of three and four color-based features.

	Scores (%)	1 color	2 colors	3 colors	4 colors	5 colors
NB	Precision	59.1	59.0	61.1	61.9	61.9
	Recall	59.1	59.0	61.1	61.9	61.9
	F-score	59.1	59.0	60.9	61.9	61.9
	Accuracy	59.1	59.0	61.1	61.9	61.9
DT	Precision	61.6	67.4	69.1	68.9	68.5
	Recall	61.3	65.7	68.8	68.7	68.3
	F-score	61.2	64.9	68.6	68.6	68.2
	Accuracy	61.3	65.7	68.8	68.7	68.1
PNN	Precision	61.3	66.2	69.1	68.0	66.6
	Recall	61.2	65.4	69.1	66.8	65.5
	F-score	61.1	65.0	69.1	66.2	65.8
	Accuracy	61.1	65.4	69.1	66.8	66.5
NB-Tree	Precision	**68.7**	**69.8**	**72.7**	**72.5**	**73.9**
	Recall	**69.7**	**68.6**	**72.8**	**72.9**	**73.8**
	F-score	**68.7**	**69.9**	**72.9**	**72.5**	**73.9**
	Accuracy	**70.7**	**71.2**	**73.3**	**73.8**	**74.2**

Table 2. Accuracy of the experiment results for dataset T3 after applying color quantization and sorting

On the whole, the data in Table 1 and Table 2 show that quantization and sorting of colors result in a significant increase in accuracy, especially when all five-color features are used with Naïve Bayes/Decision-Tree Hybrid (NB-Tree) classifiers and when three-color features are used with the Probabilistic Neural Network (PNN). In fact, these two classifiers obtain overall accuracy results of 74.2% and 69.1% with quantization and sorting. Without quantization and sorting these two classifiers achieve only 65.2% and 66.6% 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 except the case of the three-color features where it roughly ties its previous performance. In fact, the performance of this classifier drops overall with color quantization and sorting.

Figure 6 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 6 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 6.

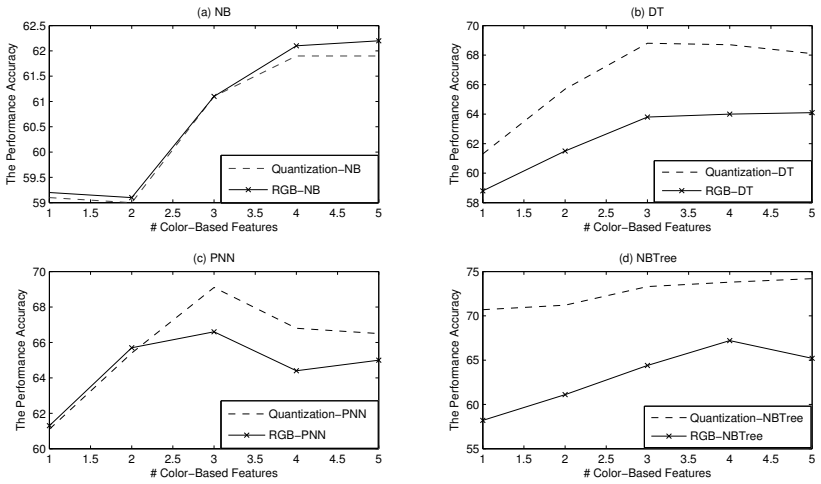

Fig. 6. Accuracy of the four classifiers on dataset T3 using different numbers of color-based features

Table 3 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 74.2% over a 50% baseline of both genders, a gain of about 24.2%.

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. [3] utilize 15.6 million features with each feature corresponding to a unique word extracted from a tweet. Similarly, Rao et al. [6] use 1.25 million features extracted from tweets.

Figure 7 shows the difference in colors chosen by female vs. male Twitter users. On the top we show popular colors chosen by female users (after clustering); the colors for male users are shown on the bottom of the figure.

Figure 8 shows the effects of different training set sizes on the accuracy of the predictions. Similar to Figure 6, 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.

All diagrams refer to data set T3. In general, the accuracy of our predictions grows linearly in the size of the training sets; larger training sets yield better accuracy results. The four classifiers exhibit similar behaviors with respect to

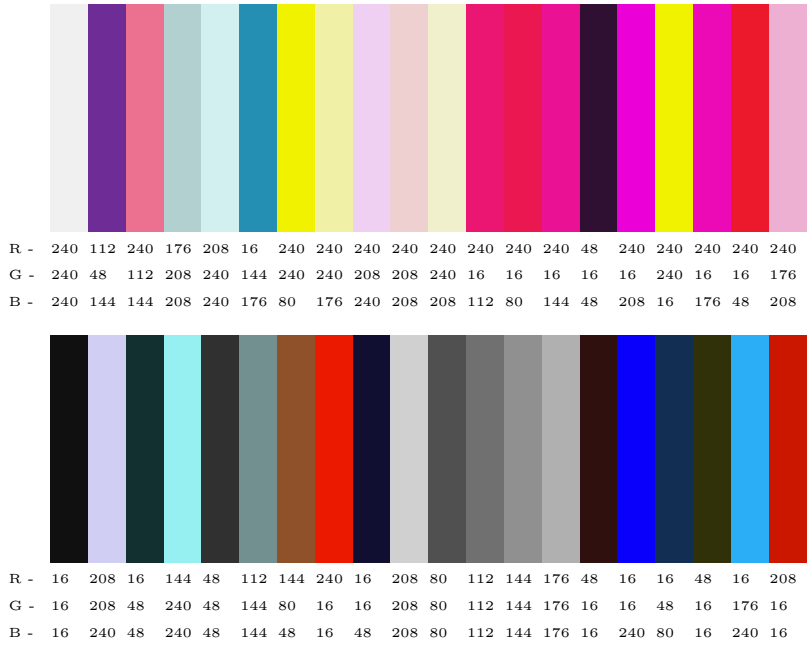

Fig. 7. Spectrum of popular colors for female users (top) and male users (bottom)

training set size. However, the performance of the classifiers differs depending on the number colors considered. In particular, the PNN classifier does best when three colors are used. Evidently, the inclusion of the sidebar fill color and border color has an adverse effect on the performance of this classifier. The DT and NB-Tree classifiers exhibit similar performance in the case of three, four and five colors. The performance of the DT classifier drops significantly when two colors are used, even more so in the case of one color. The NB-Tree classifier also exhibits a performance drop in the case of two colors and one color; however, this classifier appears to be less sensitive to the number of colors than the DT classifier. Finally, the NB classifier shows the worst performance of the four classifiers we considered; however, this classifier benefits when larger color sets (consisting of 5 and 4 colors) are used. We conclude that the NB-Tree classifier is the most suitable for our gender predictions. Not only does this classifier yield the highest accuracy results; it is also more robust than the other classifier when fewer colors are considered.

5.2 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

	Scores (%)	T1	T2	T3	T4
NB	Precision	64.1	63.0	61.9	62.7
	Recall	64.2	63.1	61.9	62.7
	F-score	64.2	63.1	61.9	62.7
	Accuracy	64.3	63.2	61.9	62.6
DT	Precision	69.3.0	69.3	68.5	61.4
	Recall	68.9	69.5	68.3	60
	F-score	69.9	69.4	68.2	60.7
	Accuracy	69.9	69.5	68.1	63.8
PNN	Precision	62.0	67.6	66.6	67.3
	Recall	61.4	65.6	63.5	64.6
	F-score	61.0	64.6	61.8	63.2
	Accuracy	61.4	65.6	63.5	64.6
NB-Tree	Precision	**72.3**	**71.6**	**73.9**	**71.9**
	Recall	**72.0**	**71.4**	**73.8**	**71.4**
	F-score	**72.1**	**71.5**	**73.9**	**71.2**
	Accuracy	**72.3**	**72.0**	**74.2**	**71.4**

Table 3. Accuracy of the experimental results for the four different datasets with color quantization and sorting

ground truth. Evidently, a complete evaluation of all 169,449 Twitter users would be impractical. We manually spot-checked about 10,000 out of the 169,449 profiles in our dataset or about 6.0% 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 169,449 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. Apparently, little will be gained by using larger datasets.

6 Conclusions and Future Work

In this paper, we studied gender classification on Twitter. We presented a novel approach for predicting gender utilizing only five color-based features extracted from the profile layout colors. Unlike existing works that use millions of features, we used only five color-based features. Despite the challenging feature-based characteristics for gender classification, we proposed color-based model for gender classification. We applied quantization colors procedure to the color-based features that compressed the color from 24-bits to 9-bits and produced discrete set of 512 colors. We empirically proved the validity of our approach by examining different classifiers over large Twitter data set collection. Our approach is using an agent with advanced colors preferences to search all profiles and predict-

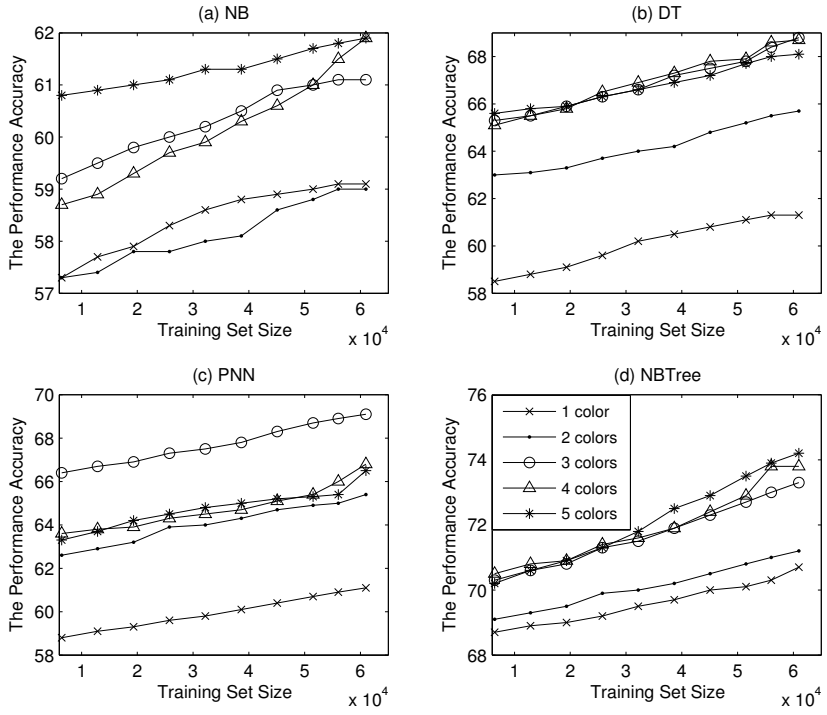

Fig. 8. Effects of different training set sizes on accuracy of different classifiers on dataset T3 with different numbers of color-based features

ing gender. Our empirical studies show that our method is reasonably accurate and highly efficient in terms of computational complexity.

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) and to incorporate them with the profile's colors.

References

1. D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani, "The Twitter of babel: Mapping world languages through microblogging platforms," *PloS one*, vol. 8, no. 4, pp. 1–9, 2013.
2. R. Wauters, "Only 50% of Twitter messages are in english, study says," http://techcrunch.com/2010/02/24/twitter-languages/.
3. J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating gender on Twitter," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, July 2011, pp. 1301–1309. [Online]. Available: http://www.aclweb.org/anthology/D11-1120

4. F. Al Zamal, W. Liu, and D. Ruths, "Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors." in *6th International AAAI Conference on Weblogs and Social Media (ICWSM'12)*, 2012.
5. W. Liu, F. Al Zamal, and D. Ruths, "Using social media to infer gender composition of commuter populations," in *Proceedings of the When the City Meets the Citizen Workshop, the International Conference on Weblogs and Social Media*, 2012.
6. D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in Twitter," in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, 2010, pp. 37–44.
7. W. Liu and D. Ruths, "Whats in a name? using first names as features for gender inference in Twitter," in *2013 AAAI Spring Symposium Series, In Symposium on Analyzing Microtext*, 2013.
8. J. S. Alowibdi, U. A. Buy, and P. S. Yu, "Empirical evaluation of profile characteristics gender classification on Twitter," in *The 12th International Conference on Machine Learning and Applications (ICMLA)*, vol. 1, Dec 2013, pp. 365–369.
9. J. S. Alowibdi, U. A. Buy, and P. S. Yu, "Language independent gender classification on Twitter," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM'13*, Aug 2013, pp. 739–743.
10. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.
11. M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel, "Knime-the konstanz information miner: version 2.0 and beyond," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 26–31, 2009.
12. S. Singh, "A pilot study on gender differences in conversational speech on lexical richness measures," *Literary and Linguistic Computing*, vol. 16, no. 3, pp. 251–264, 2001.
13. S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni, "Gender, genre, and writing style in formal written texts," *Text*, vol. 23, no. 3, pp. 321–346, 2003.
14. M. Koppel, S. Argamon, and A. R. Shimoni, "Automatically categorizing written texts by author gender," *Literary and Linguistic Computing*, vol. 17, no. 4, pp. 401–412, 2002.
15. R. Sarawgi, K. Gajulapalli, and Y. Choi, "Gender attribution: Tracing stylometric evidence beyond topic and genre," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, Portland, Oregon, USA, June 2011, pp. 78–86.
16. S. Nowson, J. Oberlander, and A. Gill, "Weblogs, genres and individual differences," in *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, Stresa, Italy, 2005, pp. 1666–1671.
17. T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat, and F. Can, "Chat mining for gender prediction," in *Advances in Information Systems*. Springer, 2006, pp. 274–283.
18. A. Mukherjee and B. Liu, "Improving gender classification of blog authors," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics, October 2010, pp. 207–217. [Online]. Available: http://www.aclweb.org/anthology/D10-1021
19. C. Peersman, W. Daelemans, and L. Van Vaerenbergh, "Predicting age and gender in online social networks," in *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, 2011, pp. 37–44.

20. S. C. Herring and J. C. Paolillo, "Gender and genre variation in weblogs," *Journal of Sociolinguistics*, vol. 10, no. 4, pp. 439–459, 2006.
21. S. Brain, "Twitter statistics," http://www.statisticbrain.com/twitter-statistics.
22. t. Business, "Who is on Twitter?" https://business.twitter.com/whos-twitter.