Feature Selection and Parameter Tuning on Age and Gender Classification for English Tweets using SVMs
A report for Automatic Classification and Kernel Methods

Roy Khristopher Bayot

Universidade de Évora, Department of Informatics,
Rua Romão Ramalho nº59, 7000-671 Évora, Portugal
d11668@alunos.uevora.pt

1 Introduction

Author profiling has been of importance in the recent years. From a forensic standpoint for example, it could be used to determine potential suspects by getting linguistic profiles and identifying characteristics. From a business intelligence perspective, companies could target specific people through online advertising. By knowing the profile of the authors, companies would easily find what a specific group of people talk about online and device strategies to advertise to these people. They could also analyze product reviews and know what types of products are liked or disliked by certain people.

Part of the reason why the interest in author profiling grows is because the growth of the internet where text is one of the main forms of communication. Through this growth, various corpora could be extracted, curated, assembled from different sources such as blogs, websites, customer reviews, and even twitter posts. Of course, this presents some problems. For example, people from different countries who use the same online platform such as Twitter or Blogger could behave differently in terms of text usage. This presents a difficulty in profiling. This work tries to take this difficulty into account by studying which kind of features are useful for different languages.

The aim of this work is to investigate the parameters for support vector machines in terms of classification using the dataset given in PAN 2015 [13]. The dataset contains twitter data from 4 different languages which are used to profile an author based on age, gender, and 5 personality traits - agreeability, conscientiousness, extrovertedness, openness, and stability. The four languages are English, Dutch, Italian, and Spanish. However, the focus of this work is on English alone and only in age and gender classification. Furthermore, the investigation is more on using different kernels and different parameters for the classification.

2 State of the Art

One of the first few works on author profiling is that of Argamon et al. in [1] where texts are categorized base on gender, age, native language, and personality. For personality, only neuroticism was checked. The corpus comes from different sources. The age and gender have the same corpus taken from blog postings. The native language corpus was taken from International Corpus of Learner English. Personality was taken from essays of pyschology students from University of Texas in Austin. Two types of features were obtained: content-based features and style-based features and Bayesian Multinomial Regression was used as a classifier. Argamon et al. had some interesting results where from the gender task, they were able to achieve 76.1% accuracy using style and content features. For age task with 3 classes, the accuracy was at 77.7% also using style and content features. For the native language task, the classifiers were able to achieve 82.3% using only content features. And finally, in checking for neuroticism, the highest obtained was 65.7% using only style features.

There has also been some research that uses datasets collected from social media. A particular example is that of Schler et al. in [15] where writing styles in blogs are related to age and gender. Stylistic and content features were extracted from 71,000 different blogs and a Multi-Class Real Winnow was used to learn the models to classify the blogs. Stylistic features included parts-of-speech tags, function words, hyperlinks, and non-dictionary words. Content features included word unigrams with high information gain. The accuracy achieved was around 80% for gender classification and 75% for age identification.

The work or Argamon et al. [1] became the basis for the work in PAN. It is an ongoing project from CLEF with author profiling as one of its tasks. It currently has three editions. In the first edition of PAN [12] in 2013, the task was age and gender profiling for English and Spanish blogs. There were a variety of methods used. One set includes content-based features such as bag of words, named entities, dictionary words, slang words, contractions, sentiment words, and emotion words. Another would be stylistic features such as frequencies, punctuations, POS, HTML use, readability measures, and other various statistics. There are also features that are n-grams based, IR-based, and collocations-based. Named entities, sentiment words, emotion words, and slang, contractions and words with character flooding were also considered. After extracting the features, the classifiers that were used were the following - decision trees, Support Vector Machines, logistic regression, Naïve Bayes, Maximum Entropy, Stochastic Gradient Descent, and random forests. The work of Lopez-Monroy in [6] was considered the winner for the task although they placed second for both English and Spanish with an accuracy of 38.13% and 41.58% respectively. They used second order representation based on relationships between documents and profiles. The work of Meina et al. [9] used collocations and placed first for English with a total accuracy of 38.94%. On the other hand, the work of Santosh et al. in [14] gave a total accuracy of 42.08% after using POS features for Spanish.

In PAN 2014 [11], the task was profiling authors with text from four different sources - social media, twitter, blogs, and hotel reviews. Most of the approaches used in this edition are similar to the previous year. In [5], the method used to represent terms in a space of profiles and then represent the documents in the space of profiles and subprofiles were built using expectation maximization clustering. This is the same method as in the previous year in [6]. In [7], ngrams were used with stopwords, punctuations, and emoticons retained, and then idf count was also used before placed into a classifier. Liblinear logistic regression returned with the best result. In [17], different features were used that were related to length (number of characters, words, sentences), information retrieval (cosine similarity, okapi BM25), and readability (Flesch-Kincaid readability, correctness, style). Another approach is to use term vector model representation as in [16]. For the work of Marquardt et al. in [8], they used a combination of content-based features (MRC, LIWC, sentiments) and stylistic features (readability, html tags, spelling and grammatical error, emoticons, total number of posts, number of capitalized letters number of capitalized words). Classifiers also varied for this edition. There was the use of logistic regression, multinomial Naïve Bayes, liblinear, random forests, Support Vector Machines, and decision tables. The method of Lopez-Monroy in [5] gave the best result with an average accuracy of 28.95% on all corpus-types and languages.

3 Dataset and Tools

The dataset for the problem at hand is composed of a set of tweets for English. Different models were made for each classification task - age and gender. There were 4 categories for the age classification - 18-24, 25-34, 35-49, and 50 and above. Gender has two categories - male and female. There were 152 users for English. Each user has different number of tweets. The dataset is balanced based on gender. Processing the data was done through Python using the scikits-learn [10] library.

4 Methodology

The focus of this study is to determine optimal parameters for classification using support vector machines. The approach is to do preprocessing, feature extraction, feature selection, then use the optimal features into support vector machines with different parameters. Evaluation was made through 10 fold cross validation.

Wilcoxon signed rank test [18] was used to compare the statistical significance of the results. This test allows the comparison between two experiments done on the same data set without making any assumptions on the distributions from which the set is drawn. A confidence interval of 95% was used. Therefore, if there was an experiment between two settings and the returned p-value after a Wilcoxon signed rank test yields less than 0.05, the null hypothesis is rejected. This means that the values are statistically different. If it yields a value greater than 0.05, it means that values between the two experiments are not statistically different.

The study will look on the polynomial kernels and radial basis function kernels. It will also explore the effect of using the information of the other class on the classification. For instance, when dealing with age classification, gender information would be used in conjunction with other features. Finally, this study will also explore the use of ensemble methods.

4.1 Preprocessing and Feature Extraction

For each language, xml files from each user are read. Then the tweets taken from each user are extracted and concatenated into one line to form one training example. The examples are then transformed by putting them all in lower case. No stop words are removed. Hashtags, numbers, mentions, shares, and retweets were not processed or transformed to anything else. They were retained as is and therefore will correspond to another item in the dictionary of features. The resulting file is used for feature extraction. Features extracted were simply *tfidf* features.

4.2 Feature Selection

After extracting *tfidf* features, we want to reduce the number of features in the text. To do that, classification was done using Support Vector Machines [3] with a linear kernel, choosing C=1 as the default. But instead of using all the *tfidf* features, different number of features were tested. The number of features were ranked by either information gain or gain ratio. The results for the accuracies by varying the feature set is given in table 1 in the succeeding chapter. However, it is suffice to say that for the succeeding experiment on polynomial kernels and radial basis function kernels, the number of features selected will be 9000 and 7000 for age and gender classification respectively. The discussion on selecting the number of features is given in the succeeding chapter.

4.3 Exploring different Kernels

Using Support Vector Machines [3] entails the use of kernels that maps the features into a different space such that separation would be done in the new space. In the earlier section, only the linear kernel was used but we further used polynomial kernels and radial basis function kernels. Polynomial kernels maps the input features to another feature space that uses the polynomial function over the similarity of the input features. Mathematically, the kernel is given by equation 1 but scikits-learn implementation has a gamma to scale the dot product as in equation 2.

$$K(x, y) = (x^T y + c)^d \tag{1}$$

$$K(x) = (\gamma(x^T x) + c)^d \tag{2}$$

For our experiments, we set the gamma to be equal to 1 but the degrees d to vary between 1, 2, and 3. For both age and gender classification, c varies between 0.0001, 0.001, 0.1, 1, 10, 1000, 10000.

Radial basis function kernel is another kernel explored. Mathematically, it is given by equation 3. It is similar the scikits implementation but with a regularization factor c.

$$K(x, x') = exp\left(-\frac{\parallel x - x' \parallel^2}{2\sigma^2}\right) \tag{3}$$

For both age and gender classification, σ and c were chosen to be one among 0.0001, 0.001, 0.1, 1, 10, 1000, and 10000.

4.4 Classification with string substitution

One of the interesting experiments is to check if the addition of some twitter specific features could affect the classification. For this experiment we only look into hashtags and links because user mentions are already handled previously since users have been anonymized. The treatment is simple and done in the preprocessing side. After taking the text from the html and converting them into lowercase, string substitution was performed on links and hashtags. All links were transformed into the text " LINK_HERE " while hashtags were turned into " HASHTAG_HERE ". The idea is that each link and hashtag wont be considered as a unique and that it would just add to the number of links or hashtags. After string substitution, *tfidf* was used and then features were ranked based on information gain. Top 9000 and 7000 words were used age and gender classification respectively. Finally, classification was done using the optimal settings given by the previous experiments and that given in the succeeding chapter.

4.5 Using Information of the Other Class as a Feature

Another interesting experiment is that of classification when using the information from the other class as a feature. For example, the results of gender classification will be used as a feature for age classification. And this will be done vice-versa with gender classification.

4.6 Exploring Ensemble Methods

Finally, a comparison with ensemble methods was also done. Two different ensemble methods were considered - Random Forests and AdaBoost. Random Forests is one of the ensemble methods wherein results are averaged in the end. The full details of the algorithm are given by the paper of Breiman in [2]. The idea is to build a multitude of decision trees and averaging their prediction results. The trees built will vary since they are built by taking a random sample with replacement from the training set. Furthermore, the features will also vary since it will be a random subset of features that are selected when the splitting is done in the training phase.

Adaptive Boosting or AdaBoost is another ensemble method formulated by Freund and Schapire in [4]. The idea is to combine different weak models to produce a powerful ensemble. It is considered adaptive since the succeeding weak learners give more weight in classifying correctly instances in the training set which was misclassified by previous weak learners.

5 Results and Discussion

5.1 Feature Selection

The results for feature selection are given in table 1. The highest accuracy could be attained when using all the different features. However, Wilcoxon signed rank test [18] was done between the experiment with the highest accuracy and other subsequent experiments within each task and each ranking function. For example, in age classification, information gain features were used and the experiment that used 26263 features was compared to the other experiment that used 10000, 9000, and so on. The numbers in boldface are those which are not statistically different from each other. Therefore, for further experiments, 9000 features were used for classification while 7000 features were used for gender classfication. Information gain ranking was used instead of gain ratio because the words given by information gain were subjectively more descriptive than those given by gain ratio.

	Age		Gender	
Num Features	info gain	gain ratio	info gain	gain ratio
26263	**0.6913**	**0.6913**	**0.6825**	**0.6825**
10000	**0.6321**	**0.6321**	**0.6167**	**0.6167**
9000	**0.6321**	**0.6321**	**0.6163**	**0.6163**
8000	0.6188	0.6188	**0.6233**	**0.6233**
7000	0.6188	0.6188	**0.6300**	**0.6300**
5000	0.6188	0.6188	0.4863	0.4863
2000	0.6188	0.6188	0.4983	0.4983
1000	0.6188	0.6188	0.4983	0.4983
700	0.6188	0.6188	0.4921	0.4921
500	0.6188	0.6188	0.4921	0.4921
300	0.6188	0.6188	0.4921	0.4921
200	0.6188	0.6125	0.4921	0.4921
100	0.6188	0.3950	0.4921	0.4921

Table 1. Accuracies that result from using different number of important features ranked by information gain and gain ratio. SVM linear kernel with C=1 was used.

5.2 Different Kernels

The results for age classification using support vector machines with polynomial kernel is given by table 2. The highest accuracy obtained was 80.92%. This was obtained from three different settings, degree 3 with C to be either 10, 1000, 10000. However, checking on Wilcoxon signed rank test, these results are not statistically significant to 80.25% given by settings with degree 2 and C to be either 10, 1000, 10000. These results are also not statistically significant against 75% with degree 3 and C to be 1. The results shown in boldface are the highest accuracies that are not statistically different from each other.

C	degree		
	1	2	3
0.0001	0.6321	0.6192	0.5121
0.001	0.6321	0.6192	0.5121
0.1	0.6321	0.6254	0.5992
1	0.6321	0.7104	**0.7500**
10	0.6321	**0.8025**	**0.8092**
1000	0.6321	**0.8025**	**0.8092**
10000	0.6321	**0.8025**	**0.8092**

Table 2. Accuracy results of using SVM with polynomial kernel on age classication task with different C and degree parameters using top 9000 informative features ranked by information gain.

The results for age classification using support vector machines with radial basis function kernel is given by table 3. Results show that the highest achieved accuracy was 80.92%. It was obtained from the kernel with gamma to be 0.001 and with a C to be 10000. This wasnt statistically different from the settings that gave 80.25%, 69.71%, 68.46%, and 65.17% which are all shown in boldface.

C	gamma						
	0.0001	0.001	0.1	1	10	1000	10000
0.0001	0.3950	0.3950	0.3950	0.3950	0.3950	0.3950	0.3950
0.001	0.3950	0.3950	0.3950	0.3950	0.3950	0.3950	0.3950
0.1	0.3950	0.3950	0.3950	0.4746	0.3950	0.3950	0.3950
1	0.3950	0.3950	0.6054	**0.6517**	0.5933	0.3950	0.3950
10	0.3950	0.3950	**0.6846**	**0.8025**	0.6196	0.3950	0.3950
1000	0.6188	**0.6971**	**0.8025**	**0.8025**	0.6196	0.3950	0.3950
10000	0.6971	**0.8092**	**0.8025**	**0.8025**	0.6196	0.3950	0.3950

Table 3. Accuracy results of using SVM with radial basis function kernel for age classification with different C and gamma parameters using top 9000 informative features ranked by information gain.

Comparing the two different kernels for age classification including different parameters, the settings settled for would be with the polynomial kernel of degree 3 with C equal to 10.

The results for gender classification using support vector machines with polynomial function kernel is given by table 4. Results show that the highest achieved accuracy was 79.54%. This comes from six different settings. The first three were when the degree is 2 and C was chosen to be either 10, 1000, or 10000. The last three settings were when degree is 3 with C also either 10, 1000, or 10000. These results are not statistically different from that which gave 70.25% which are all written in boldface.

C	degree		
	1	2	3
0.0001	0.6300	0.4983	0.4733
0.001	0.6300	0.4983	0.4733
0.1	0.6300	0.5233	0.4733
1	0.6300	0.6367	**0.7025**
10	0.6300	**0.7954**	**0.7954**
1000	0.6300	**0.7954**	**0.7954**
10000	0.6300	**0.7954**	**0.7954**

Table 4. Accuracy results of using SVM with polynomial kernel for gender classification with different C and degree parameters using the top 7000 informative features ranked by information gain.

The results for gender classification using support vector machines with radial basis function kernel is given by table 5. Results show that the highest achieved accuracy was 80.79%. This comes from the settings with gamma to be 1 and C to be 10. However, these arent statistically different from settings which gave 80.21%, 80.13%, and 79.54% accuracy, all of which are written in boldface in the table.

Comparing the two different kernels for gender classification including different parameters, the settings settled for would be with the radial basis function kernel with gamma equal to 1 and C equal to 10.

5.3 String Substitution

Comparisons between the the best from previous parameter tuning experiments and that where links and hashtags were substituted with a different string tag is shown in table 6. We can see that using the same parameters for the classifier, the performance drops when there's string substitution but the results are not statistically different.

C	gamma						
	0.0001	0.001	0.1	1	10	1000	10000
0.0001	0.5233	0.5233	0.5233	0.5171	0.4921	0.5171	0.4733
0.001	0.5233	0.5233	0.5233	0.5171	0.4921	0.5171	0.4733
0.1	0.5233	0.5233	0.5233	0.5171	0.4921	0.5171	0.4733
1	0.5233	0.5233	0.5233	0.6167	0.6629	0.4796	0.4733
10	0.5233	0.5233	0.6238	**0.8079**	0.7025	0.4796	0.4733
1000	0.5233	0.6367	**0.8021**	**0.8013**	0.7025	0.4796	0.4733
10000	0.6367	**0.7954**	**0.8021**	**0.8013**	0.7025	0.4796	0.4733

Table 5. Accuracy results of using SVM with radial basis function kernel for gender classification with different C and gamma parameters using the top 7000 informative features ranked by information gain.

Task	Accuracy			Settings			
	With String Sub	Best from Previous Experiments	P-value	kernel	gamma	degree	C
Age	0.7892	0.8092	0.8206	poly	N/A	3	10
Gender	0.7442	0.8079	0.4963	rbf	1	N/A	10

Table 6. Comparison between the best accuracy for age and gender classification from previous experiments against a classifier trained with links and hashtags string substitution.

5.4 Using Information from Other Class

Comparisons between the the best from previous parameter tuning experiments and another method where the training set features is augmented by the information of the other class is given by table **??**. We can see that age classification with features augmented by gender classification results has 2% increased accuracy but it is not statistically different. For gender classification, the results are almost the same and are also not statistically different.

Task	Accuracy			Settings			
	With String Sub	Best from Previous Experiments	P-value	kernel	gamma	degree	C
Age	0.8292	0.8092	0.5453	poly	N/A	3	10
Gender	0.8021	0.8079	0.8206	rbf	1	N/A	10

Table 7. Comparison between the best accuracy for age and gender classification from previous experiments against a classifier trained with a feature set added with the information of the other class.

5.5 Ensemble Methods

Accuracy results for random forests are given in tables 8 and 9. Accuracy results for AdaBoost are given in table 10 and 11.

n-estimator	Accuracy
10	0.6058
100	0.6046
1000	0.6112
2000	0.6046
5000	0.6108
10000	0.6175

Table 8. Accuracy results for age classification using forests of randomized trees.

n-estimator	Accuracy
10	0.5983
100	0.6500
1000	0.7029
2000	0.6762
5000	0.6963
10000	0.6967

Table 9. Accuracy results for gender classification using forests of randomized trees.

n-estimator	Accuracy
50	0.5479
100	0.5217
150	0.5475
200	0.5146
250	0.5275

Table 10. Accuracy results for age classification using AdaBoost.

n-estimator	Accuracy
50	0.6787
100	0.7187
150	0.7183
200	0.7500
250	0.7308

Table 11. My caption

6 Conclusions and Future Work

References

1. Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123, 2009.
2. Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
3. Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
4. Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.
5. A Pastor López-Monroy, Manuel Montes-y Gómez, Hugo Jair Escalante, and Luis Villaseñor-Pineda. Using intra-profile information for author profiling.
6. Adrian Pastor Lopez-Monroy, Manuel Montes-y Gomez, Hugo Jair Escalante, Luis Villasenor-Pineda, and Esaú Villatoro-Tello. Inaoe's participation at pan'13: Author profiling task. In *CLEF 2013 Evaluation Labs and Workshop*, 2013.
7. Suraj Maharjan, Prasha Shrestha, and Thamar Solorio. A simple approach to author profiling in mapreduce.
8. James Marquardt, Golnoosh Farnadi, Gayathri Vasudevan, Marie-Francine Moens, Sergio Davalos, Ankur Teredesai, and Martine De Cock. Age and gender identification in social media. *Proceedings of CLEF 2014 Evaluation Labs*, 2014.
9. Michał Meina, Karolina Brodzinska, Bartosz Celmer, Maja Czoków, Martyna Patera, Jakub Pezacki, and Mateusz Wilk. Ensemble-based classification for author profiling using various features. *Notebook Papers of CLEF*, 2013.
10. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
11. Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. Overview of the 2nd author profiling task at pan 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum (Working Notes)*, 2014.
12. Francisco Rangel, Paolo Rosso, Moshe Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. Overview of the author profiling task at pan 2013. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, pages 352–365. CELCT, 2013.
13. Francisco Rangel, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. Overview of the 3rd Author Profiling Task at PAN 2015. In L Cappellato, N Ferro, J Gareth, and E San Juan, editors, *CLEF 2015 Labs and Workshops, Notebook Papers*. CEUR-WS.org, 2015.
14. K Santosh, Romil Bansal, Mihir Shekhar, and Vasudeva Varma. Author profiling: Predicting age and gender from blogs. *Notebook Papers of CLEF*, 2013.
15. Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205, 2006.
16. Julio Villena-Román and José Carlos González-Cristóbal. Daedalus at pan 2014: Guessing tweet author's gender and age.

17. Edson RD Weren, Viviane P Moreira, and José PM de Oliveira. Exploring information retrieval features for author profiling—notebook for pan at clef 2014. *Cappellato et al.[6].*
18. Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, pages 80–83, 1945.