Age and Gender Classification using Glove Vectors and Support Vector Machines

Roy Khristopher Bayot

Universidade de Évora, Department of Informatics,
Rua Romão Ramalho n°59, 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. Bayesian Multinomial Regression was used because it was shown to be effective for text classification problems. It's a variant of logistic regression that is used instead of naive bayes classifiers because it doesnt assume independence between features.

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. The winnow algorithm is a supervised learning algorithm that learns a linear classifier. It is similar to a perceptron, that updates the weight in every training example. The difference lies in the fact that for perceptron, the weight update is additive while for Winnow, the update is multiplicative. Considering scale, Multi-Class Real Winnow becomes much more efficient than SVM. A possible drawback could happen when the decision bounderies are non-linear.

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. Most of the submissions for this edition used deci-

sion trees, wherein a classification tree is learned by splitting the data based on information gain or gini index of the features from which the split is based on. The idea is to keep on splitting until the instances in the leaves come from only one class. Three used Support Vector Machines wherein the features are mapped into another space and a hyperplane is fitted on the new space. The fitted hyperplane is made such that the gap between the different classes in the new space is as wide as possible. Two approaches used logistic regression, where it measures the relationship between the category and it's features by estimating the probabilities using a logistic function. This is then used to predict. One used Naïve Bayes where a probabilistic classifier is constructed using Bayes' theorem but with independence assumptions of features. Another used Maximum Entropy, which is the multivariate form of logistic regression. Another used Stochastic Gradient Descent, which is an iterative method for sum-minimizations such as that in least squares. And finally, one used Random Forest, which is an ensemble method described later in the text, where a multitude of trees is used instead of using one tree for classification.

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. They used liblinear but did not specify which classifier was used.

The work of Meina et al. [9] used collocations and placed first for English with a total accuracy of 38.94%. They experimented on different classifiers but Random Forests gave the best result. Their final Random Forest classifier was trained on a 12-core machine with 30GB of RAM, and parameters were obtained through trial and error. The parameters include minimum samples per leaf to be equal to 5, the size of the feature set for each tree is equal to $\sqrt{n\_features}$, and the number of trees was lowered to 666 although it converges at values higher than 1000.

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. They used three different kinds of features - content based, style based, and topic based. Each of the features has a classifier. Content based features in the form of ngrams and style based features in the form of ngrams of POS tags both used Support Vector Machines. Although the kernel and parameters used were not mentioned. Topic based features in the form of LDA topic model was used as features for a Maximum Entropy classifier. The results from each of the three classifiers are fed finally into a decision tree.

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 re-

turned 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.

INSERT TEXTS ABOUT WORD2VEC, GLOVE, AND CLASSIFICATION PAPERS USING SUCH.

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. More data statistics could be found in table 1.

	min	max	average	std	median
number of tweets of each user	32	100	93.20	16.82	100
total length of tweets of each user	1979	12485	7445.30	2389.61	7438.5
average tweet length of each user	29.46	124.85	79.61	20.28	80.35

Table 1. Various statistics on twitter data.

Looking at the number of tweets per user, we can see that the minimum is 32 tweets, the maximum is 100 tweets per user, with an average of 93.20, a standard deviation of 16.82, and a median of 100. Looking at the total length of all tweets for each user, the minimum is 1979 characters, the maximum is 12485 characters, an average of 7445.30 with a standard devation of 2389.61, and a median of 7438.5. Finally, looking into the average tweet length per user, the minimum is 29.46, the maximum is 124.85, average is 79.61, with a standard deviation of 20.28, and a median of 80.35.

Processing the data was done through Python using the scikits-learn [10] library. The GloVe pretrained model was taken from **INSERT HERE WHERE GLOVE WAS TAKEN** which was then loaded into gensim **INSERT GENSIM REF HERE**. GloVe has many different pretrained models varying from

25, 50, 100, and 200 dimensions. These models were pretrained using 2 billion tweets.

4 Methodology

This study has three main purposes. First, it is to evaluate age and gender classification of twitter text with minimal preprocessing using GloVe word vectors. It will check for different dimensions of word vectors. Second, it is to evaluate if using word vector averages or word vector centroids, in varying dimensions would yield better classification results. Finally, it will check if adding more preprocessing into the pipeline would yield better classification results.

The overview approach is to do preprocessing, feature creation through word vector lookup and averaging or centroid, and then using the created features into support vector machines. Evaluation was made through 10 fold cross validation.

4.1 Preprocessing

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.

4.2 Feature Creation

After minimal preprocessing, feature creation is done by utilizing pretrained GloVe vectors. The GloVe vectors are loaded into a gensim word2vec model. For each line that forms one training example, the words are looked up from the model. If the word does not exist in the model's dictionary, it is discarded. If the word does exist ins the model's dictionary, the vector is stored and the processed further. There are two ways that the vector is processed. The first is that for each training example, the average word vector is taken. The second is that the centroid is taken. The resulting vector is then used as the features for training that will be fed to the support vector machine. All these are done in the four available dimensions given by the pretrained GloVe model - 25, 50, 100, 200.

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.

5 Results and Discussion

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.