Age and Gender Identification using Stacking for Classification*
Notebook for PAN at CLEF 2016

Madhulika Agrawal and Teresa Gonçalves

Universidade de Évora, Portugal
madhu1agrawal@gmail.com,tcg@uevora.pt

Abstract. This paper presents our approach of identifying the profile of an unknown user based on the activities of known users. The aim of author profiling task of PAN@CLEF 2016 is cross-genre identification of the gender and age of an unknown user. This means training the system using the behavior of different users from one social media platform and identifying the profile of other user on some different platform. Instead of using single classifier to build the system we used a combination of different classifiers, also known as stacking. This approach allowed us explore the strength of all the classifiers and minimize the bias or error enforced by a single classifier.

1 Introduction

In the modern age of technology where everybody likes to be well connected with the world, the social media platforms provides an excellent opportunity to do so. They also provide a way to express one's views openly. By using Facebook, Twitter, Whatspp, Snapchat and various other applications, we create a huge collection of information belonging to different genres. Along with the personal details that is shared through these applications, a massive chunk of data that is dependent on the profile and personality of the user also gets generated.

Every individual have a different style of writing. The structure of sentence, vocabulary and way of representation of thoughts varies from person to person. In spite of these differences, people belonging to similar group share certain aspects of writing. By similar group we mean people belonging to same gender, or same age group or same geographic location among few. Thus when people transcribe their thoughts into posts on social media platforms they contribute to the pool of data of the group to which they belong. This data can be further used for developing many other systems.

The author profiling task of PAN@CLEF 2016 [4], focuses on this aspect of human communication. The question here is that, given the activities of a set of users from one genre, is it possible to identify the profile of another user

* This system is submitted as practical work done for one of the Ph.D courses on Automatic Classification and Kernel Methods, at Universidade de Évora.

which belong to some other genre. Different platforms provides different facilities and at the same time impose various restriction on the way of representation of thoughts. Now it will be interesting to see if the writing style changes with the change in platform or not.

The rest of this paper is organized as follows: Section 2 gives a description about the dataset used. Section 3 presents the approach of feature selection and combining of classifiers. Section 4 describes the results obtained on training data during the development as well the results on test dataset. Section 6 concludes the paper.

2 Dataset Description

The dataset is part of Author Profiling task of PAN@CLEF 2016 [4]. The dataset consist of xml documents containing tweets from various users. Each dataset corresponds to documents in one of the languages: English, Spanish or Dutch. The dataset have documents written by user belonging to age groups 18-24, 25-34, 35-49, 50-64, 65-xx. Users are also classified according to their genders, male and female. Number of documents belonging to each age group is different. But the number of documents by male and female are same across all the three datasets. A detailed description of the dataset is given in Tab.1. The Dutch dataset do not have age details.

Table 1. Number of documents of each category in author profiling training dataset 2016

Category	Dutch	English	Spanish
Gender			
Male	192	218	125
Female	192	218	125
Age Group (in years)			
18-24	-	28	16
25-34	-	140	64
35-49	-	182	126
50-64	-	80	38
65-xx	-	6	6
Total	384	436	250

3 Experiments

In our experiments, we re-framed the problem of identifying the gender and age of the author as a classification problem. The classifier is trained over the given classes (male and female for gender and different age groups for the age). Then

the idea is to classify the new document as belonging to one of these classes. The system was trained separately for gender and age classification. Instead of using a single classifier we used a combination of classifiers, also known as stacking. Our experiment can be categorized into following steps:

- Preprocessing
- Feature Extraction and Feature Selection
- Classification

3.1 Preprocessing

All the tweets from a single user are joined together into one document. Once we have the documents for all the users, we perform few preprocessing steps to remove the noise from the corpus. All the HTML/XML tags from the documents were removed. Any reference to other user was replaced with @USERNAME. @LINKS were used to represent links to other web pages. If there are some emotions expressed in the tweet using an emoticons, they were replaced with @EMOJI. Duplicate tweets, extra whitespace, tabs and blank lines were also removed. Whole text was converted to lower case and the stop words were removed.

3.2 Feature Extraction and Feature Selection

The documents are then represented as *TF-IDF* matrix [5]. This TF-IDF conversion resulted in the feature space with much higher dimension. Many features in this feature space does not contribute in the classification and hence it is reduced by evaluating the worth of a feature by measuring the information gain with respect to the class [7]. The information gained by each term for identifying the categories is calculated and the terms having certain threshold value of information gain are retained. The threshold parameter was set to 0 in our experiments, meaning all the attributes for which the information gain is positive are retained. The percentage reduction in the feature space is represented in the Tab.2.

Table 2. The percentage reduction in the feature space based on information gain

Dataset	Original Dimension	Reduced Dimension		% Reduction	
		Gender	Age	Gender	Age
English	38267	979	122	97.44	99.68
Spanish	32587	503	70	98.45	99.78
Dutch	14180	358	-	97.47	-

3.3 Classification

Stacking [6] is an ensemble learning method where several hypotheses are combined into one. It consist of base models and a meta model as shown in fig.1. Each base model is an individual classifier with their own hypothesis. The classification decision made by each of these base classifiers are feed as input to the meta classifier, which is responsible for making the final classification decision. The tool used by us for performing the classification is Weka [2], developed by University of Waikato. The Tab.3 shows the base and meta classifiers used by our approach for gender and age classification.

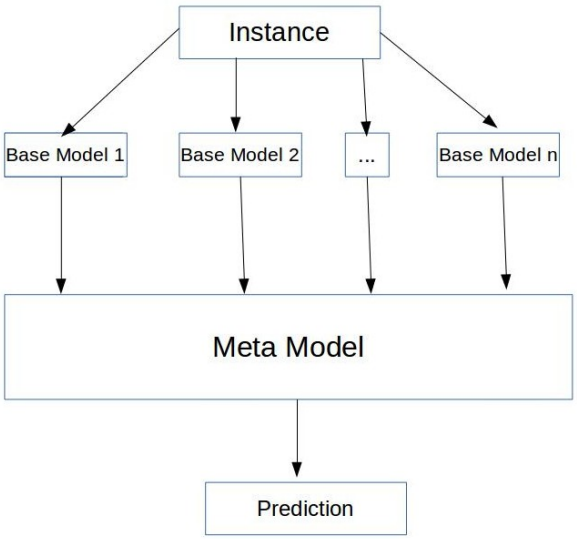

Fig. 1. Stacking

Table 3. Base and meta classifiers used for gender and age classification

Category	Base Classifier	Meta Classifier
Gender	Bayesian Logistic Regression Naive Bayes Multinomial Naive Bayes Linear SVM	Naive Bayes
Age	Naive Bayes Multinomial Simple Logistics Naive Bayes Linear SVM	Linear SVM

4 Results

During the development, the accuracy obtained for gender and age identification in all the three languages using 10-fold cross-validation is given in the Tab.4. The results obtained after submitting the developed system on the virtual machine TIRA [1, 3] and running it on the test datasets are as shown in Tab.5. The tests were conducted on two datasets, test1 and test2. Both these datasets were collected from reviews in case of Dutch. Concretely test1 is 10% of test2. For English and Spanish, test1 was collected from social media and test2 from blogs. The important observation from the results of both the development and the test dataset is that the performance of the system is consistent throughout the languages for gender classification. Thus the classifier that is used in our approach performs same irrespective of the language of dataset. This is a key take away from this experiment as it is important to be able to develop a system that can identify the user, irrespective of the its language. .

Table 4. Accuracy of classifying gender and age during development using 10-fold cross-validation

Dataset	Gender (%)	Age (%)
English	96.10	64.22
Spanish	96.4	66.8
Dutch	94.01	-

Table 5. Accuracy of classifying gender and age on various test datasets.

Dataset	Accuracy		
	Gender	Age	Overall
test1-Dutch	0.5000	-	-
test1-English	0.5000	0.2586	0.1207
test1-Spanish	0.4688	0.2500	0.1094
test2-Dutch	0.5080	-	-
test2-English	0.5128	0.3846	0.1923
test2-Spanish	0.5357	0.4821	0.2857

The performance of our system for test1 in Spanish perform worse than the baseline that choses always the most frequent class, for gender classification (<50%).

5 Conclusion

In this paper, we have discussed the performance of combining several classifiers into one. The results that was obtained on the test dataset are poor as compared

to the results obtained during the development of the system. This shows that the cross-genre author profiling is a challenging task. It would be interesting to see if by using some other features, the performance can be improved or not.

The future work may include fine tuning of certain parameters such as the threshold for determining the information gain. If we can identify attributes that contributes most to the classification than it might improve the system's performance.

References

1. Gollub, T., Stein, B., Burrows, S., Hoppe, D.: TIRA: Configuring, Executing, and Disseminating Information Retrieval Experiments. In: Tjoa, A., Liddle, S., Schewe, K.D., Zhou, X. (eds.) 9th International Workshop on Text-based Information Retrieval (TIR 12) at DEXA. pp. 151–155. IEEE, Los Alamitos, California (Sep 2012)
2. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD explorations newsletter 11(1), 10–18 (2009)
3. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14). pp. 268–299. Springer, Berlin Heidelberg New York (Sep 2014)
4. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th Author Profiling Task at PAN 2016: Cross-genre Evaluations. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2016)
5. Salton, G., Fox, E.A., Wu, H.: Extended boolean information retrieval. Communications of the ACM 26(11), 1022–1036 (1983)
6. Wolpert, D.H.: Stacked generalization. Neural networks 5(2), 241–259 (1992)
7. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: ICML. vol. 97, pp. 412–420 (1997)