Age and Gender Identification using Stacking for Classification

Madhulika Agrawal, Teresa Gonçalves
madhu1agrawal@gmail.com,tcg@uevora.pt

Abstract

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 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.

Data

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.

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

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

Results

During the development, 10-fold cross-validation was used for calculating the accuracy for gender and age identification in all the three languages. The tests were conducted on two datasets, test1 and test2. For Dutch, both these datasets were collected from reviews. Concretely test1 is 10% of test2. For English and Spanish, test1 was collected from social media and test2 from blogs.

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

Table 2: Accuracy of classifying gender and age during development using 10-fold cross-validation.

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

Table 3: Accuracy of classifying gender and age on various test datasets.

Our Approach

The problem of identifying the gender and age of the author is re-framed 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 single classifier we used a combination of classifiers, also known as stacking. Our experiment can be categorized into following steps:

- **Pre-processing:**
  - Combine all the tweets from a user into a single document.
  - Remove HTML/XML tags from the documents.
  - Replace user references with USERNAME, links to other web pages as Links and emoticons as EMOJI.
- **Feature Extraction & Feature Selection:**
  - Represent the documents as TF-IDF matrix.
  - The information gained by each term for identifying the classes is calculated and the terms having positive information gain are retained.

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	-

Table 4: The percentage reduction in the feature space based on information gain.

- **Classification:** Stacking consists of few base models and a meta model. Each base model is an individual classifier with their own hypothesis. The classification decision made by each of these base classifiers are fed as input to the meta classifier, which is responsible for making the final classification decision.

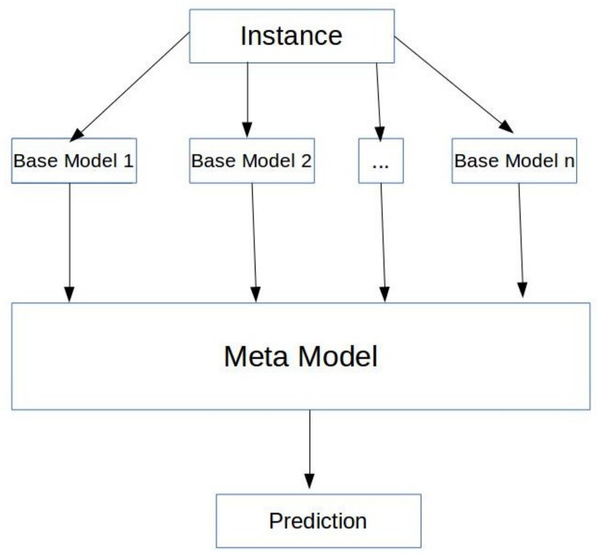

Figure 1: Stacking

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

Table 5: Base and meta classifiers used for gender and age classification.

References

[1] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[2] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[3] Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 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, September 2016.