Genre-Agnostic Features For Author Profiling

Pepa Gencheva[1] Martin Boyanov[1] Preslav Nakov[2] Yasen Kiprov[1] Ivan Koychev[1] Georgi Georgiev[1]

FMI, Sofia University "St. Kliment Ohridski", Sofia, Bulgaria {mkbojanov, pkgencheva, koychev}@uni-sofia.bg, g.d.georgiev@gmail.com

Qatar Computing Research Institute, HBKU, Doha, Qatar pnakov@qf.org.qa

Abstract

We present our solution for the PAN-2016 Author Profiling Task. The task was to predict the gender and the age group of a person given several samples of his/her writing, and it was offered for three different languages: English, Spanish, and Dutch. Our approach focused on extracting genre-agnostic features such as *bag-of-words*, *sentiment and topic derivation*, and *stylistic features*. We then used these features to train SVM-based classifiers for gender and age-group and adapted to the target domain via self-training.

Introduction

Author Profiling is a task in Natural Language Processing that aims at identifying various characteristics of the authors by analyzing texts written by them. The task can range from classifying the author by his/her age, gender or mother tongue, to finding his/her socio-economic category.

The PAN-2016 Author Profiling Task asks participants to identify the gender and age-group of a person, given a set of documents s/he has authored. The task is even more challenging, because the system is given training data only for social media documents, but the evaluation is performed on data from another genre. Furthermore, the task is held in English, Spanish and Dutch. Thus, the participants must provide a cross-genre multi-lingual solution to the problem.

Our main focus is on extracting *genre- and language-agnostic features* based on the content of the documents written by the author. We also experiment with *bootstrapping* the algorithm via several iterations of learning and classification.

Pipepeline

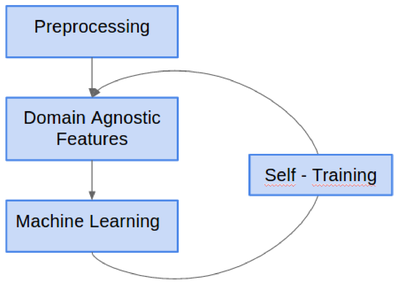

Figure 1: System pipeline

Important Result

Domain-agnostic features perform well in the general case. Further improvements can be achieved by *adapting* to the target domain via *self-training*.

Features

We concentrated on genre-agnostic features which can always be computed regardless of the media. The features we implemented reflected the writing style of the author as well as the semantic purpose of the documents.

Function Words	Out of Dictionary	Unique words
Type - Token ratio	Capital letters	Average word length
Pointwise Mutual Information	Sentiment	Profanity

Table 1: Features

Stereotypes

The distribution of the features allowed us to confirm or contradict some stereotypes about various differences between the genders and the age groups. For example, we found that on average *young ladies swear the most* and that the word 'sex' was most indicative for the 50-64 age group.

The youngest age group would talk about their parents and paying the bills and the 25-34s were more concerned with starting a family.

Also, men tended to talk about sports, computer games and beer, whereas women would discuss makeup and clothes.

Results

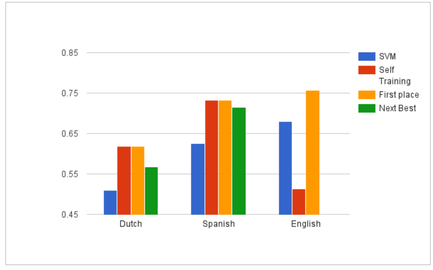

Figure 2: Results for test-dataset-2 Gender classification. They show that *self-training is a viable approach* - we achieved *first place for Dutch and Spanish gender classification*. However, it also showed that the *approach was unstable* as it *decreased our score for English*.

Self-Training

We employed a technique called *self-training* to *adapt to the target genre*:

1. Run the pipeline on the training data
2. Use the classifier to classify the test data
3. Add the documents classified with high confidence back into the training set
4. Go back to 1 or quit after a certain number of iterations

The technique showed some promising results, but we found it hard to tune. It introduces two new parameters - the number of iterations and the confidence threshold - and we speculate that they are different for every (language, genre) pair.

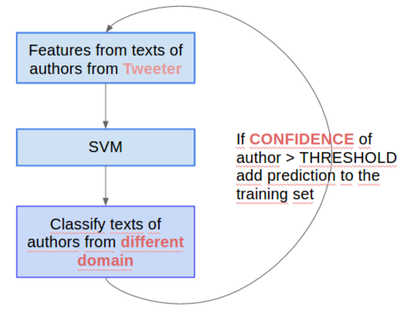

Figure 3: Self Training Approach

Conclusion

Author profiling is an interesting problem and it is quite challenging in the domain agnostic setup. We have produced a framework which can be easily extended with new features or another learning method. Self training can be used to adapt to the target domain, but it is hard to tune and the parameters could be different for every (domain,language) pair.