PANcakes Team: A Composite System of Genre-Agnostic Features For Author Profiling

Notebook for PAN at CLEF 2016

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

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

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

Abstract We present the system we built for participating in the PAN-2016 Author Profiling Task [9]. The task asked 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. We participated in both subtasks, for all three languages. 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, as implemented in LIBLINEAR for the gender classification sub-task, and in LIBSVM for the age classification sub-task.

1 Introduction

Author Profiling is a task in Natural Language Processing that aims at identifying different 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 [9] 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 would be performed over data in 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. Their systems are evaluated through TIRA [5], which is a platform for evaluation as a service.

In this paper, we present our approach to the Author Profiling Task. Our main focus is on extracting *genre- and language-agnostic features* based on the content of the documents written by the author. We also experimented with bootstrapping the algorithm via several iterations of learning and classification.

The paper is structured as follows. In the next section, we give a brief overview of some related work. Then, in the following sections, we describe in detail the data used, the different steps we undertake for tackling the task, as well as the evaluation of the system.

2 Related Work

The task of Author Profiling was first introduced to the PAN series of scientific events in 2013. It started as a task for identifying the author's gender and age in English and Spanish. In the following years, the notion of posts' genre was introduced, the Dutch language was added to the task, and in 2015 the task also included author personality profiling.

The participants from the previous years applied a large variety of techniques. They used different pre-processing steps like removing the HTML code, the hash-tags or user mentions from Tweeter posts, replacing or removing URLs, etc. The features used represent a wide variety from simple style-based features to more complex content-based features and combinations thereof [6].

In terms of cross-genre classification, this is the first year that the PAN Author Profiling Task is setup to evaluate the algorithms on domains different than the training domain.

3 Data

For training we used the dataset provided by the PAN-2016 Author Profiling Task organizers. It consisted of three different training sets for the different languages: English, Spanish, and Dutch. The English, Spanish, and Dutch training datasets contained Twitter posts for 436, 250, and 384 authors, respectively. For the purpose of the task, only the posts' text could be used for training.

The training datasets contained labels, classifying the authors by gender (male or female), and by age-group (18-24, 25-34, 35-49, 50-64, 65+). For Dutch, only the gender subtask was available.

For testing, we used training sets from the PAN-2014 Author Profiling Task, which contained posts in three non-Twitter domains: social, blogs, and reviews. These datasets are available for English and Spanish only. For Dutch, we used the training dataset from the PAN-2015 Author-Profiling Task, even though it is from the Twitter domain. We also assembled our own test dataset, which contained posts from the three available non-Twitter domains from the PAN-2014 Author Profiling Task. It was also used for testing.

We also used the NRC Word-Emotion Association Lexicon [7]. It contains 14,182 English words associated with positive and negative sentiment and emotions such as anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. These words are translated into several other languages, so that the associations could be used in them, too. We use the Lexicon for English, Spanish, and Dutch. Another type of Lexicons we used were lists of swear words for English, Spanish, and Dutch, which were accumulated from several resources [10,11,12].

4 Our Genre-Agnostic Approach

We designed and implemented a system that could be easily configured with different parameters and extended with new features.

We then proceeded to extract a large variety of features. They can be divided into sentiment and lexical features. The lexical features can further be split into content and stylistic features. The semantic features and the content-based lexical features are based on the pre-processed tokens, while the style-based lexical features are based both on the pre-processed and on the raw authors' posts. From these features, we further selected those that serve best in the classification task and we used them in the training of our model. We concentrated on training the model for age and gender classification in English. For the models for Dutch and Spanish, we used the same features and model parameters as the ones chosen for English.

The system is built on top of the tools, provided by the scikit-learn machine learning library in Python [8]. It contains various machine learning algorithms, some feature extraction mechanisms and a handy way to create a pipeline of all the features and the chosen classification method.

In the next subsections, we describe in detail the individual steps in the model construction.

4.1 Pre-processing

The pre-processing step is an important part of the domain-agnostic author profiling task. For this step, we remove all the genre-specific strings from the authors' posts. Most of them are Twitter-style sequences. We also join all the author's posts to create bigger text and eliminate the differences in the posts' length in the different domains such as Tweeter and blogs.

We first clean the posts from HTML tags. Then, we remove all genre-specific sequences in the text including strings such as user-mentions, at-mentions, hash-tags, URLs, punctuation sequences, emoticons, etc. The next step is to tokenize the cleaned text. The tokens created in this way are used for most of the features. Some of the features use the raw post as well.

4.2 Features

Lexical Style-based Features

The lexical style-based features aim at finding a correlation between author's gender and age and his/her manner of writing with the use of simple *surface metrics*. We count the number of the following: function words used by the author; words that are not from a predefined dictionary of words for each of the languages; words starting with a capital letter; words written with all capital letters; the total number of sentences; punctuation signs in the post's text; URLs used in the posts; e-mails used in the post; phone numbers used; different pronouns used. We also include the type-token ratio, which counts the number of unique words used by the author divided by the total number of words used, the average word length, the average length of a sentence, the numbers used in the post. We further take into account the number of img and br HTML tags used in the post.

Another feature we use is finding whether the person has mentioned something that makes it obvious whether s/he is a man or woman, e.g., “my wife”, “my man”, “my girlfriend”, “my boyfriend”, etc. We do this for English only. We devise such a feature for the author’s age too, searching for the pattern “I am ”, followed by a number, which is supposed to signify the age of the author. We also count the frequencies of Part-of-speech tags. For this purpose, we employ the Natural Language Toolkit for Python.[1] We also tried to include some readability metrics such as the Automatic Readability Index, but they showed poor performance on the test sets.

From these features, we selected the best-performing ones. For age classification, we eventually used the number of function words used, the number of img HTML tags, the number of sentences, the ratio of capital letters over all letters, and the number of punctuation signs used. For the gender classification task, we included the strings, showing that the author obviously belongs to one of the classes, the average length of the words, the number of img HTML tags, the proportion of capital letters over all letters, and the number of punctuation signs used. All of these features are included in the models for English, Spanish and Dutch.

Lexical Content-based Features

The content-based features look deeper into the words used and the topics in the posts of the authors. The feature that was the most essential one was *bag-of-words*. For age classification, we transformed the authors’ posts into a matrix of word uni-gram and bi-gram counts. For the gender classification task, we employ a similar bag-of-words approach and further transform the count matrix into a normalized TF-IDF representation. The document frequency we use is logarithmic and the normalization is cosine. For gender, classification TF-IDF bag-of words approach proved to be effective with char tri-grams and word uni- and bi-grams. We tuned this feature to use only tokens with document frequency between 30 and 80 percent. Another interesting content feature is the Non-negative Matrix Factorization (NMF), which we use for topic extraction. We extracted a total of 20 topics. This feature is used in both age and gender classification.

We created a dictionaries containing information about the *Point-wise Mutual Information* (PMI) between the words and the classes. These dictionaries were extracted from the training data for every language. During the testing phase, we calculate the sum of the PMI for all the words for a given author and provide them as features for the classifier. Unfortunately, this approach did not perform very well in the cross-domain genre because of the different vocabulary one uses when switching to another context.

Semantic Features

The semantic features look into the overall sentiment expressed by the words, used in the posts. For this purpose, we use the NRC Word-Emotion Association Lexicon [7]. We accumulate the sentiment from each of the words. The sentiment is collected for positive and negative meanings of the words, as well as for the emotions of anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. From these sentiment meanings, we finally chose the best-performing ones, which are the following: positive, negative, joy, surprise, and trust. They showed good performance for the age classification for all of the three languages and were not included in the gender classification.

For gender classification, we also used the number of swear words used. For the different languages, we collected lists of swear words from several resources [10,11,12].

5 Experiments & Evaluation

For building the classification models for the two tasks and the three languages, we tried a large variety of machine learning algorithms and different ensembles thereof. For gender classification, liblinear proved to perform better, while for age classification, libsvm with a radial basis kernel worked best. We used the implementations of the algorithms from scikit-learn, which are built on top of LIBLINEAR [4] and LIBSVM [2]. For the final models, we tuned the parameters for liblinear (C was set at 0.7), and for libsvm (C was set to 1.25 and gamma was set to 0.125). From all of the features, we selected the best-performing ones on the different test sets, described in the third section. In the following tables, we represent the scores we achieved for the available domains and for the real test set (testset2 on Tira). In Table 1, we present the accuracy for the author gender classification, and in Table 2, we present the accuracy for the author age-group classification.

5.1 Feature Selection

For feature selection, we used several approaches. We tried setting a variance threshold of 80 percent, meaning that all of the features that have 0/1 value for more that 80 percents of the training examples will be removed. We also tried selecting different percentages of the features based on the F-value between label/feature for classification tasks or the chi-square statistics of non-negative features for classification tasks. Unfortunately, none of the feature selection attempts yielded good results on the test sets and we did not use feature selection for the final models.

5.2 Self Training

We figured out that if we could classify a couple of authors from the target domain with high confidence, then we could use their corresponding data for training. This approach is known as self training and shows good results in a number of NLP tasks [3]. Thus, we modified the system to pass through several iterations of training and testing. At each iteration, the entries classified with high confidence by the SVM classifier would be given as training data for the next iteration. We hoped that this way we could provide more of the needed data from the target domain. And indeed, this was the case in some of our experiments. However, this approach introduces two new parameters to the system: the confidence level to accept new training data and the number of iterations. They were hard to tune for the different test sets and if not tuned well, the results were worse. Thus, we did not include it in out last submission, because we could not be sure how they would perform on the unknown test set.

	blogs	reviews	social	test
English	0.7823	0.5747	0.5371	0.6795
Spanish	0.7840	0.5668	-	0.6250
Dutch	-	-	-	0.51

Table 1. Accuracy of gender classification results for different languages.

	blogs	reviews	social	test
English	0.4013	0.23581	0.2728	0.3718
Spanish	0.4772	0.25471	-	0.2679

Table 2. Accuracy of age-group classification results for English and Spanish.

6 Conclusion

We have presented the system developed by our team for participating in PAN-2016 Author Profiling Task. It included different lexical and semantic features and used liblinear and libsvm for training the model for the age and gender classification tasks in the three languages. The system is easily extendable and can serve as a basis for other attempts to tackle the problem.

For future improvement and investigation of the problem of domain-agnostic author profiling, we find the bootstrapping approach interesting to develop and work on. It showed some promising results on the official test set, where for Dutch we got an accuracy score of above 0.70 for the gender task. However, it performed really poorly on the other languages and it was not part of our final submitted system. We believe that this approach can lead to better results, but further study is required to identify the correct parameters.

Acknowledgments

This research was performed by a team of students from MSc programs in Computer Science in the Sofia University “St Kliment Ohridski”.

We thank the Sofia University “St Kliment Ohridski” for the support and guidance to our team participation at the CLEF 2016 Conference.

References

1. Bird, Steven, E.L., Klein, E.: Natural Language Processing with Python. O'Reilly Media Inc. (2009)
2. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011)
3. Chapelle, O., Schlkopf, B., Zien, A.: Semi-supervised learning (2010)
4. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research 9, 1871–1874 (2008)
5. Gollub, T., Stein, B., Burrows, S., Hoppe, D.: TIRA: Configuring, Executing, and Disseminating Information Retrieval Experiments. In: Proceedings of the 9th International Workshop on Text-based Information Retrieval (TIR 12) at DEXA. pp. 151–155. Los Alamitos, California (2012)
6. Grivas, A., Krithara, A., Giannakopoulos, G.: Author profiling using stylometric and structural feature groupings : Notebook for PAN at CLEF 2015 (2015)
7. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word-emotion association lexicon. Computational Intelligence 29(3), 436–465 (2013)
8. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
9. 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)
10. Wikipedia: Dutch profanity - wikipedia, the free encyclopedia (2016), https://en.wikipedia.org/w/index.php?title=Dutch_profanity&oldid=719768946, [Online; accessed 18-May-2016]
11. Wikipedia: English profanity - wikipedia, the free encyclopedia (2016), https://en.wiktionary.org/wiki/Category:English_vulgarities, [Online; accessed 18-May-2016]
12. Wikipedia: Spanish vulgarities - wikipedia, the free encyclopedia (2016), https://en.wiktionary.org/wiki/Category:Spanish_vulgarities, [Online; accessed 18-May-2016]