# Using Wide Range of Features for Author profiling

## Notebook for PAN at CLEF 2015

Suraj Maharjan and Thamar Solorio

University of Houston
Department of Computer Science,
Houston, TX, 77004
smaharjan2@uh.edu, solorio@cs.uh.edu

**Abstract** Predicting an author's age, gender and personality traits by analyzing his/her documents is important in forensics, marketing and resolving authorship disputes. Our system combines different styles, lexicons, topics, familial tokens and different categories of character n-grams as features to build a logistic regression model for four different languages: English, Spanish, Italian and Dutch. With this model, we obtained global ranking scores of 0.6623, 0.6547, 0.7411, 0.7662 for English, Spanish, Italian and Dutch languages respectively.

## 1 Introduction

The PAN15 [4] author profiling shared task is to predict the age-group (18-24, 25-34, 35-49, 50-xx), gender (male, female) and personality traits (extroverted, stable, agreeable, conscientious, open) of authors by analyzing their tweets. Task participants have access to the training data in four different languages: English, Spanish, Italian and Dutch.

Researchers have approached this problem in a variety of ways on a number of different datasets. In order to capture the profile of Facebook users, Schwartz et al. used what they call an Open Vocabulary approach [7]. They used word n-grams and Latent Dirichlet Allocation (LDA) topics as features and compared their method with a Closed Vocabulary approach that used Linguistic Inquiry and Word Count (LIWC) word categories and found the Open Vocabulary method to be better for all of the personality traits as well as for age and gender.

Estival et al. [1] performed the same task on emails collected from both native and non-native English speakers. Apart from age, gender and personality traits, they also tried to predict the native language, country of residence and level of education of the authors. They experimented with different classifiers such as SMO, Random Forest and SVM and found that for different attributes of an author's profile. They also tried feature selection and while using all of their character, lexical and structural features worked best for age and gender prediction, removing the lexical features produced better results for personality traits.

In this paper, we built two separate models for age-gender and personality for each language. We used a wide range of features as described in Section 2 for profiling author's age, gender and personality traits. We experimented with different combinations of these features with Logistic Regression as a classifier.

## 2 Methodology

We started out by tokenizing author's twitter data with the help of Ark Tweet NLP tokenizer [2]. This tool is well adapted for Twitter. In addition, we replaced all of the hyperlinks/urls with *URL* . Also, we expanded most of the contractions used in Twitter. For instance, we replaced *r* with *are*, *u* with *you*, *n't* with *not*. We replaced good and bad emoticons with words *emoticon_good* and *emoticon_bad* respectively. After preprocessing and tokenization, we extracted the following features:

**Lexical:** These consist of the word unigrams, bigrams and trigrams, which are commonly used in an author's profile.

**Twitter Style:** Stylistic features capture the stereotypical style of a particular group of authors. Number of words, characters, question marks, exclamation marks, hashtags, user mentions (@), urls, all capitalized words, text and number combined tokens, average word length and average tweet length features were used to capture the writing style of authors.

**Familial Tokens:** Some of the familial tokens used by males and females are very distinct. Females are a lot more likely to use *my hubby, my bf, my husband, my boyfriend, etc.* than males. In the same way males will use words such as *my wife, my girlfriend, my gf, etc*. The presence of these tokens in authors' tweets are strong indicative features to distinguish the gender of an author. We prepared a list of these tokens for all languages and used their normalized counts as features.

**Categorical Char n-grams:** Sapkota et al. [6] evaluated the predictive power of different subgroups of character n-grams in single and cross domain authorship attribution settings. They defined ten different character n-gram categories based on affixes, words and punctuation. Instead of using all the character n-grams, we defined similar categories and used those combinations of categories of char n-grams that gave us high accuracy. Categories like *mid-word, beg-punct, multi-word, prefix, mid-punct and space-prefix* proved to be good for this task.

**LDA Topics:** Many researchers have used LDA topics as features in order to predict gender, age group and personality of authors. We also used similarity of tweets with LDA topics as feature. For age and gender we clustered documents into eight topic groups and for personality we clustered into ten topic groups.

**Age and Gender:** We hypothesized that the same personality trait might have different patterns in authors from different age groups and gender. So, we used these as features in the determination of the personality of authors.

Apart from these features, we also experimented with the word categories in the LIWC corpus. However, the addition of LIWC word lists degraded our system's performance and we dropped the LIWC word list from our final system. After obtaining these features, we trained a multiclass logistic regression classifier with them. We used the *gensim* [5] Python library for LDA topic extraction and the *scikit-learn* [3] framework for feature extraction and to perform classification.

## 3 Experiments and Results

Since, the training dataset is small, we performed our experiments through ten fold cross validation. We experimented with different combinations of the above defined features. Table 1 shows the combination of features that gave us the best results for age_gender and personality task. We created two separate models for age_gender and for personality classification. The personality model uses the age_gender model's age_group and gender prediction as features.

**Table 1.** Results.

| Features | English | Spanish | Italian | Dutch |
|---|---|---|---|---|
| Lexical | ✓ | ✓ | ✓ | ✓ |
| Twitter Style | ✓ | ✓ | ✗ | ✓ |
| Familial Tokens | ✗ | ✗ | ✗ | ✓ |
| LDA Topics | ✓ | ✓ | ✗ | ✗ |
| Categorical Char n-grams | ✓ | ✓ | ✓ | ✓ |
| Age and Gender | ✓ | ✓ | ✓ | ✓ |

Table 2 shows the results on the actual PAN15 test dataset for the four different languages. The global ranking formula as defined by PAN organizers is defined in Equation 1. The $RMSE$ is the Root Mean Square Errors normalized between 0 and 1. The $joint\_accuracy$ is the combined accuracy of age and gender classification. The global ranking scores show that our system works best for Dutch and worst for Spanish language.

$$global\_ranking = (1 - RMSE) * joint\_accuracy \tag{1}$$

**Table 2.** Results on actual test dataset.

| Language | Global | RMSE | Age | Agreeable | Both | Conscientious | Extroverted | Gender | Open | Stable |
|---|---|---|---|---|---|---|---|---|---|---|
| English | 0.6623 | 0.2388 | 0.6901 | 0.2127 | 0.5634 | 0.2222 | 0.2299 | 0.7465 | 0.2645 | 0.2647 |
| Spanish | 0.6547 | 0.2702 | 0.625 | 0.2569 | 0.5795 | 0.2357 | 0.3008 | 0.7955 | 0.2696 | 0.288 |
| Italian | 0.7411 | 0.2122 | | 0.2118 | | 0.2225 | 0.161 | 0.6944 | 0.2476 | 0.2181 |
| Dutch | 0.7662 | 0.2488 | | 0.2781 | | 0.2378 | 0.2102 | 0.7813 | 0.2358 | 0.2821 |

## 4 Discussion and Conclusion

Our system combined various style, lexicons, topics, familial tokens and categories of character n-grams features to build a final logistic regression classifier. Building a model with all of these features combined did not give good performance. But when combined selectively, these features boosted our system's performance. In addition, we also found that the same type of features are not the ones that are strongly representative of authors' profiles across different languages. For instance, the categories of character n-grams that were prominent across different languages were not the same. For English, Spanish and Italian, familial tokens did not improve the system performance, whereas for Dutch it is one of the key features. Similarly, Twitter specific style features are prominent across English, Spanish and Dutch but not in Italian. However, across all languages, word unigrams, bigrams and trigrams are important features, which illustrates that authors having similar attributes tend to use similar words. In conclusion, our system analyzes a wide range of features to profile author's age-group, gender and personality traits and is reasonably successful in doing so.

## Acknowledgments

We thank PAN15 organizers and committee members for organizing Author Profiling task. This research was funded by National Science Foundation CAREER grant 1462141.

## References

1. Estival, D., Gaustad, T., Pham, S.B., Radford, W., Hutchinson, B.: Author profiling for English Emails. In: Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics. pp. 263–272 (2007)
2. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: Part-of-speech tagging for twitter: Annotation, features, and experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2. pp. 42–47. HLT '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2002736.2002747
3. 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)
4. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at PAN 2015. In: Cappellato L., Ferro N., Gareth J. and San Juan E. (Eds). (Eds.) CLEF 2015 Labs and Workshops, Notebook Papers. CEUR-WS.org vol. 1391, (2015)
5. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), http://is.muni.cz/publication/884893/en

6. Sapkota, U., Bethard, S., Montes, M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 93–102. Association for Computational Linguistics, Denver, Colorado (May–June 2015), http://www.aclweb.org/anthology/N15-1010
7. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E., et al.: Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one 8(9), e73791 (2013)