GronUP: Groningen User Profiling
Notebook for PAN at CLEF 2016

Mart Busger op Vollenbroek, Talvany Carlotto, Tim Kreutz, Maria Medvedeva,
Chris Pool, Johannes Bjerva, Hessel Haagsma, and Malvina Nissim

University of Groningen, Groningen, The Netherlands
{m.j.h.busger.op.vollenbroek, t.carlotto, t.j.kreutz, m.medvedeva, c.pool.1} @student.rug.nl,
{j.bjerva, hessel.haagsma, m.nissim} @rug.nl

Abstract We train an SVM linear model on tweets to perform user profiling, in terms of gender and age, on non-Twitter social media data, whose actual nature is unknown to us at developing time. We choose features that we deem appropriate to profile authors on social media in general, and which do not characterise the specifics of Twitter data too closely. Additionally, we pay specific attention to engineering features that work beyond the actual values observed on training data, and which should thus be better at making predictions on data from a different genre. Our system works on English, Dutch, and Spanish data without any language-specific features or parameter tuning. Results from cross-validation on the training set seem to indicate that all features contribute rather equally to the model's performance, so all features are included in our final test-time model. Results on test data are lower than what is normally observed for this task — that is, when done in a single-genre setting — but are in fact the state-of-the-art for the cross-genre setting at PAN 2016.

1 Introduction

The huge volume of user generated data on social media makes it appealing to have systems that can profile users based on their production. This has obvious advantages in terms of commercial exploitation, e.g. in targeted marketing and advertising, as well as in the forensic and security areas. It is not surprising then that the author profiling tasks organised within PAN in the last years have attracted such large numbers of participants, with numerous systems being developed and evaluated [9,8,7]. One little explored aspect of profiling, though, is how portable such systems are when used in a cross-genre setting. When models are trained and tested on the same kind of user-generated data, system developers can rely on very specific aspects of that dataset or genre. A prime example is Twitter, where metadata, including hashtags, can be exploited to this end. A system that relies on such platform-specific features might not perform well when profiling users' productions on different social media.

The author profiling task at PAN 2016 is tailored to address exactly this point [10]. Indeed, while training data is provided in the form of user generated tweets, the test data is only known to be some kind of non-Twitter social media text, with no other information available at system development time regarding its nature. Participating in this task therefore means building a system that predicts gender and age of social media users who write in English, Spanish, or Dutch, exploiting features that might be typical of social media texts in general, not just Twitter content. This setting naturally prompts researchers to focus on features related to general language characteristics rather than metadata or genre-specifics.

In this paper we describe a system that fits these requirements. We discuss our approach to the problem, modelling choices, and an overview of the features we have used. When presenting results, we highlight the contribution of different feature groups to the model's overall performance. We also discuss performance by comparing our system's results with those obtained by the other participants, and by comparing our own results on the two different test sets provided by the organisers as they were drawn from different sources.

2 Approach and Data Representation

In this section we describe how the data is represented, what preprocessing is performed, and which classifier we use.

2.1 Data and preprocessing

The training data consists of tweets of users in Dutch (384 users), English (436 users) and Spanish (250 users). Gender must be predicted for all three languages, while age information is provided for English and Spanish only. Both predictions are structured as classification tasks, where gender is a two class problem, and age values are binned into five separate classes.

While gender is completely balanced with a 50/50 distribution of male and female users in all sets, the age groups are not so evenly distributed, as is shown by the number of users per age class in Table 1. The profile of a user is defined by his or her output as a whole. So, a Twitter user is represented by the collection of their tweets, which are provided in XML files, one per user (the amount of tweets per user varies, with an average of 658 for English and 856 for Spanish, while each Dutch user is represented by exactly 200 tweets). Similarly, in the test data, a user is represented by all of their produced texts, regardless of the texts' origin.

Table 1. Distribution of users per age group in the training data.

Age	English	Spanish
18-24	28	16
25-34	140	64
35-49	182	126
50-64	80	38
65-xx	6	6
Total	436	250

The first module in our system preprocesses the data. The XML files containing the tweets are converted into a data structure where each post is an item and the unique user_id is used to identify the user. In the case of training data we have a list of tweets per user. For test data, in order to convert what we assumed to be longer texts to the same structure, we split the texts into sentences and add them to the list, resulting in a list per user with one sentence per item. Next, the items (tweets/sentences) are tokenised at the word level.

Tokenisation is performed using the Natural Language Toolkit (NLTK, [2]) TweetTokenizer and the test data is split using the default NLTK word_tokenizer. Before tokenisation some tokens are converted to placeholders: all URLs are replaced with the string ‘URL’, and all numbers are replaced by ‘NUMBER’. Additionally, HTML markup is deleted in this step. In order to speed up the training and development phase, we store all tokenised documents so that they can be accessed directly at feature extraction stage.

2.2 Learning Settings

For developing our system we used the Python Scikit-learn machine learning library [5]. Building on previous work and insights from participating systems in previous editions of this task, we opted to train a Support Vector Machine (SVM) with a linear kernel. We shortly evaluated the effect of different parameter values on performance using cross-validation on training data, but as a general approach we did not want to tune the system on tweets too closely. Indeed, while we observed that increasing C led to an improvement in performance, we suspected that allowing for fewer incorrect classifications in the training data would lead to overfitting and worse performance on data from a different genre. We therefore focused on feature engineering rather than SVM settings, and used default parameters for all our models.

3 Features

In choosing features to model user traits, we relied both on previous work and on intuition. From a language perspective, we have aimed at creating general models, without features that are tailored to a specific language only. In this section and in the next we describe all of the features we implemented, while in Section 5 we include information on their specific contribution to our models’ performance.

3.1 N-grams

N-gram-based features have proven to be highly useful indicators of various linguistic differences between authors [8,1]. At the token level, features with unigrams, bigrams and trigrams were implemented. The character-level features consisted of n-grams of 2 to 5 characters. Unigrams, bigrams, and trigrams of parts-of-speech were also used (see also Section 3.11).

3.2 Starts with capital

Users do not always start a sentence with a capital letter, and we believe there is a difference between age groups in this respect. We expect younger users do be less keen on using sentence-initial capitals. We calculate this feature by taking the percentage of sentences/tweets which start with a capital letter.

3.3 Capitalised Tokens

Similarly, capitalisation of individual tokens (e.g. proper nouns) is not always done in social media writing. Implementing a script to check the correct application of capitalisation rules in a given language would involve identifying, for instance, proper nouns in the text, which would be too time-consuming for our purposes. Hence, we simply implemented this feature by calculating the proportion of capitalised words in a sentence w.r.t. the total number of words in that sentence. The final feature is the average of that proportion across items for a given user.

3.4 Capital Letters

Another capitalisation phenomenon in social media is writing in all uppercase or all lowercase letters. Although many people interpret writing in all capital case as shouting, many people still adopt it as a personal style. Although there is no conventional interpretation of what it means to write in all lower case, it is still used by some. This phenomenon is represented by the proportion of capital letters in a sentence to the total number of characters in that sentence. The resulting feature is the average of this proportion for a given user, across items.

3.5 Ends with punctuation

Users do not always end a sentence with proper punctuation, and we suspect that proper punctuation can also be used as a differentiating feature between age groups. We represent this feature as the percentage of items which have ".", "!" or "?" as their final token.

3.6 Punctuation by Sentence

Similarly, we want to capture overall punctuation use, which ranges from no punctuation at all to long sequences of "!" or "?". We count the number of common punctuation characters (namely, ",", ".", ";", "!", "?" and "-") in every sentence. Then we calculate the proportion of these characters w.r.t. the total number of characters in the sentence. Then, counts are averaged across a user's items for the final feature representation.

3.7 Average word length and average sentence length

The use of longer words and sentences can be an indicator of a more advanced writing style, and it has been observed that older users tend to use longer words and sentences. We represent this characteristic as the average word length in characters and the average sentence/tweet length in tokens.

3.8 Out of Dictionary Words

We hypothesise that the number of typos and slang words can be a useful feature for distinguishing different age groups. For this feature we calculate the percentage of misspelled and slang words out of the total number of words per user. To detect out of dictionary words in English and Spanish we use the Python Enchant Library[1] [6] with `aspell` dictionaries for all available dialects. For Dutch, the extraction of this feature proved unstable for technical reasons, and we eventually decided not to use it in the final system.

3.9 Vocabulary Richness

The level of variety of a person's language use, represented here by the richness of their vocabulary, is also a salient characteristic of their writing style.

A way to measure the vocabulary richness of a person is to count how many words used by the user were only used once. The more unique words written by a user, the richer their vocabulary. We represent vocabulary richness as the percentage of unique words across all sentences/tweets for a given user.

3.10 Function Words

Function words are understood to be an important feature for determining the gender and age of an author [11]. Here, we use function words denote the words specific to a certain category of authors. In other words, they are the words that are more frequently used in one category relative to the rest of the categories. Examples include sports-related words for male users versus female users, or popular textisms for the youngest category of uses versus the older users.

$$rf_{tc} = \frac{tf_{tc}}{tf_{t!c}}$$

The relative frequency (rf) of a term was calculated as in the equation on the right, where t denotes the term and c the class. We then used a ranked list of the top 2500 most frequent words for each category. The function word feature itself is then a vector containing the occurrences of function words for each category in a document.

3.11 Parts of Speech

The use of specific parts of speech (POS) can be a strong indication of the characteristics of a person. It has been noted, for example, that female writers on social media tend to use more pronouns while male writers use more articles and prepositions, independently of their age group [11]. For extracting POS information, we first experimented with the Stanford tagger using the English and Spanish models [13], but we eventually settled on the TNT tagger [3] trained on the Penn Treebank for English and on the CoNLL 2002 data for Spanish and Dutch. The extracted parts of speech were also used as n-gram features to grasp more structured trends (see Section 3.1).

[1] http://abisource.com/projects/enchant

3.12 Emoticons

The amount and type of emoticons used by people on Twitter can be a useful source of information when estimating the age of the user. It has been found that, on American Twitter, younger people are more likely to use emoticons without noses (e.g. ':)') than older people, who prefer emoticons with noses (e.g. ':-)') [12]. Also, there is a correlation between the overall frequency of emoticon use, usage of non-nose emoticons, and age. As such, we implement features capturing the proportion of emoticons with noses out of all emoticons, and the proportion of emoticon tokens out of all tokens.

In addition, based on our own intuitions about emoticon use among different user groups, we add features capturing the percentage of reverse emoticons (e.g. '(:') out of total emoticons, and the proportion of happy emoticons (any emoticon with ')', ']', 'P', 'D' as mouth) out of total emoticons. Emoticons are simply defined as any multi-character token starting with ':', '=', or ';'.

4 Second Order Attributes

Given the nature of this shared task, in which the most challenging aspect is to train a model able to generalise profile predictions to data from an unknown social media platform, we allowed for different configurations when representing features. For example, sentence length should really be represented as the *relative verbosity of one class to the next*, since tweets are not usually written as full sentences but other social media can be, and absolute values might thus be meaningless. The same goes for features such as punctuation, use of capital letters and other features that would not be able to say something general about gender or age, if we would use their absolute counts.

We allow for second order attribute representation of features to deal with this problem. Our approach is based on the approach in [1]. For the features that return real values, a mean is calculated for the training data, and the relative distance of the classes (male and female for gender) is determined. When transforming the test data, the scores for a user are compared to the mean for the testing data, and the difference is then compared to those found for the training data. If, in training, it is found that female users use five percent less capitals than the mean, and an unseen author get the same relative score compared to the mean for the testing data, the feature is represented as a vector in which prototypically the distance to classes ['MALE', 'FEMALE'] is [10, 0].

We also keep the normal (first order) configuration in place for features that do not benefit from a second order representation. This was the case for the function word feature and word, character and POS-tag n-grams.

5 Results and Discussion

We first discuss results on the cross-validated training set, including an assessment of feature contribution via ablation experiments (Section 5.1), before presenting the results on test data in Section 5.2.

5.1 Results on training data

Accuracy scores on the cross-validated training set are given in Table 2. These scores are obtained using the complete feature set.

Overall, while we expect results on non-tweet test data (see Section 5.2) to be lower, we observe that the results on the training data are also substantially below the state-of-the-art results for all three languages. At PAN 2015, systems—trained and tested on Twitter—achieved scores $> 85\%$ (English) and $> 95\%$ (Spanish and Dutch) for gender, and around and well above 80% for age in Spanish and English, respectively [7]. We believe that our lower scores are likely due to the conscious choice to avoid the use of potentially strong age/gender indicators which would work on Twitter only. In other words: we choose to be less accurate at profiling Twitter users, in order to perform better on the non-Twitter data.

Table 2. Accuracy scores on the cross-validated training set. The value for *both* is calculated as the percentage of cases where age and gender of a given user are both predicted correctly.

Language	Age	Gender	Both
Dutch	–	0.7213	–
English	0.4573	0.7067	0.3302
Spanish	0.4899	0.7085	0.4008

To gain a better insight into which features work best, we experimented with feature ablation and single-feature experiments. However, instead of using each single feature separately, we grouped features in coherent subsets (for example, all emoticon-related features were grouped in a subset called *emoticon*), and experimented with those. The following feature subsets have been used:

- *emoticon*
- *length*
- *grammatical correctness*
- *count (punctuation, capital tokens and capital letters per sentence)*
- *n-grams (word and character)*
- *vocabulary (richness and function words)*
- *pos-tags*

Each of the different subsets was tested individually and promising subsets were combined to see their combined effect. We report the results of these experiments in Figures 1 to 6. All of the results reported are based on five-fold cross-validation, and the scores for English and Spanish correspond to the accuracy of the joint prediction of age and gender.

For English none of the features really affected the system's performance in any direction. While the combined scores are highest when using only n-gram or vocabulary feature subsets, adding the other feature subsets does not harm the system. The Spanish results show that emoticon features are more useful there than when used for English or Dutch. The other feature subsets, again, neither harm nor improve the system. For Dutch, only the gender label was available, so combined accuracy scores could not be computed. Overall, as is known from previous work, it seems that n-gram information is crucial in profiling.

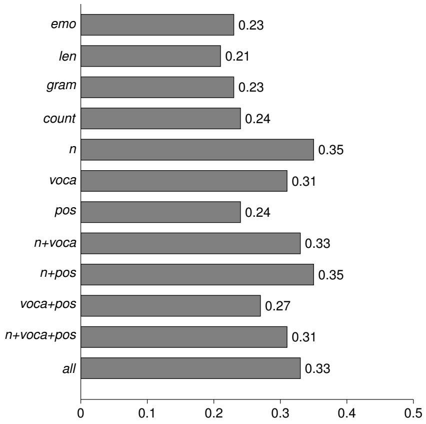

Figure 1. English accuracy results on joint age and gender prediction using individual subsets and their combinations.

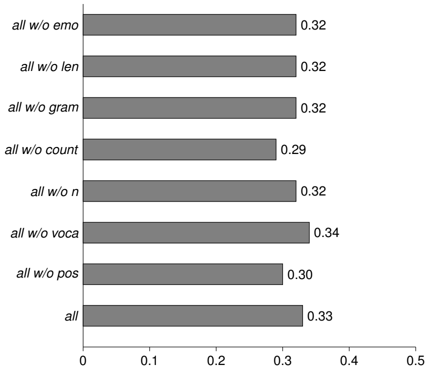

Figure 2. English accuracy results on joint age and gender prediction leaving subsets out.

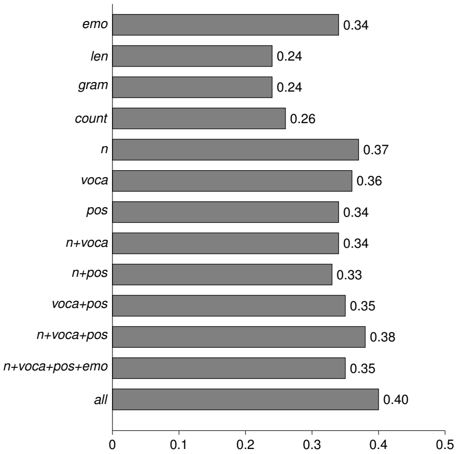

Figure 3. Spanish accuracy results on joint age and gender prediction using individual subsets and their combinations.

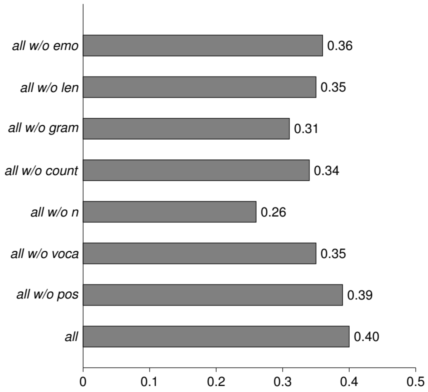

Figure 4. Spanish accuracy results on joint age and gender prediction leaving subsets out.

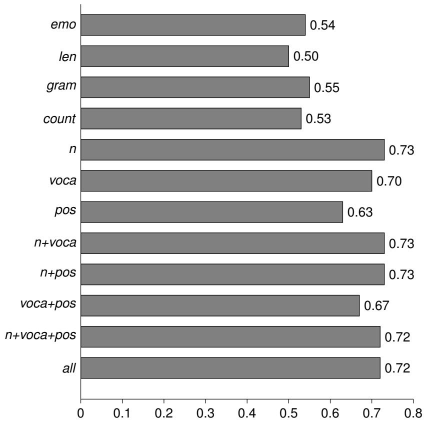

Figure 5. Dutch accuracy results on gender prediction using individual subsets and their combinations.

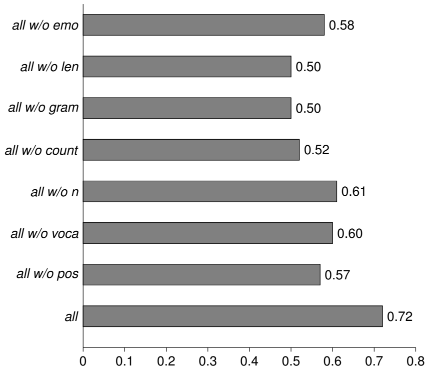

Figure 6. Dutch accuracy results on gender prediction leaving subsets out.

First Order vs. Second Order Attributes One of the main aims in developing our system was making features portable from tweets to other types of social media. This was the rationale behind the implementation of second order attributes, as opposed to 'regular' first order attributes, as explained in Section 4. To have an approximate idea of their contribution, we ran the system using either first or second order attributes only.[2] Results, obtained through five-fold cross validation on the training set, are reported in Table 3.

Table 3. Cross-validated results on training data using either first or second order attributes only.

	First Order			**Second Order**		
	Age	**Gender**	**Both**	**Age**	**Gender**	**Both**
Dutch	–	0.56	–	–	0.58	–
English	0.46	0.49	0.25	0.42	0.56	0.23
Spanish	0.49	0.66	0.34	0.51	0.63	0.33

As we can see, there are some differences in scores, but no clear trends, when using only first or only second order attributes. However, we developed second order features in order to better cope with data from different sources, bypassing the limitations caused by using absolute values. By cross-validating on the training data, we do not actually put such features to use, as the data in training and testing is all from Twitter. It is therefore left to future work with specific runs on the test data, once it will become available, to determine the actual contribution of second order attributes in cross-genre user profiling.

[2] For a proper comparison, when using the first order feature set, we only include the features that are affected by the second order treatment, while the others are left out (see Section 4).

5.2 Results on test data

Because none of the feature subsets strongly affect the system performance, we use a model incorporating all of them for the final run on the test data.

On the TIRA interface [4], the organisers posted two different test sets (test1 and test2), and asked participants to test systems on both. While we knew that test1 had been used for early development by some teams, at the time of running the system we did not know anything about the nature of the two sets. At the time of writing though, we know that for English and Spanish, test1 was collected from social media and test2 from blogs, while for Dutch both sets were collected from reviews, with test1 simply being a subset (10%) of test2.

The results we got on both test data sets are reported in Table 4. In what follows we first discuss results on the test2 sets, which are the official evaluation sets used for ranking, and later we write about the difference in performance on the two test sets.

Table 4. Accuracy scores on PAN 16 test sets. The value for *both* is calculated as the percentage of cases where age and gender of a given user are both predicted correctly.

Dataset	**Age**	**Gender**	**Both**
Dutch Test 1	–	0.5000	–
Dutch Test 2	–	0.4960	–
English Test 1	0.3046	0.5575	0.1897
English Test 2	0.5897	0.6410	0.3846
Spanish Test 1	0.2813	0.5313	0.1875
Spanish Test 2	0.5179	0.7143	0.4286

Results on Test2 As already observed for training data (see above), results on test data are generally far below typical values for this task, especially for gender. However, when comparing our results with those obtained by the other submissions [10], so as to have an idea of the difficulty of the task, we observe that we perform well.

Our system achieves the highest score for age on both languages out of all participating systems, with an accuracy of 0.5179 for Spanish and 0.5897 for English, respectively. On gender, however, performance is worse: apart from Spanish, where we achieve second best results with an accuracy of 0.7143, our score for English is 0.6410, which places us 7th in the final rank, and on Dutch data we achieve a baseline score. While we are not able to provide any informed motivation for this performance, we can observe that other systems that performed very well on the English and Spanish data, performed worse on Dutch gender as well, and vice-versa. The amount of training data for Dutch was also substantially smaller than for English and Spanish (see Section 2.1), which might have had some impact as well.

Lastly, regarding speed, we observe that our system is definitely not fast, as we generally rank second- or third-slowest w.r.t. run time.

Test1 vs Test2 Apart from Dutch, where we perform the same on both sets and indeed test1 is just a subset of test2, our system is substantially worse on test1 compared to test2 (see Table 4). Without having access to the data itself, we can only speculate on the difficulties implied by running a system trained on tweets, though without Twitter-specific features, on social media data vs. blog data, especially since 'blog' and 'non-Twitter social media' are quite broad categories.

From the results, it appears that in terms of how age and gender surface and can be captured, Twitter is more similar to blogs than other social media data, which is rather counterintuitive. It would also be most distant from reviews. However, one could also hypothesise that, in general, demographics detection on social media other than Twitter is more difficult than in blogs, but that is speculation. Similarly, observing that the performance on Dutch gender is globally much worse than on English and Spanish, can we hypothesise that authors of reviews are harder to profile, in general?

We can glean some insights into this by looking at the results for the author profiling task at PAN 2014, where the organisers provided four different sets in English and Spanish, one per genre: Twitter, social media, blogs, and reviews (the latter for English only) [8]. Training and testing was done *in-genre*, so that differences in performance across sets are more likely due to intrinsic characteristics of the genre rather than portability issues. Average results in English are highest for Twitter, followed by blogs, reviews, and social media, and in Spanish for Twitter, followed by blogs, and social media.[3] This might indeed be an indication that surface cues for predicting demographics are harder to grasp in social media (other than Twitter) than in blogs and Twitter, and that reviews are also hard for user profiling, per se. However, obtaining a more accurate picture of the reasons behind the difference in performance will require further investigation, taking specifically into account the implications of cross-genre modelling.

In any case, our system shows respectable portability, considering that even on test1 it achieves the highest score of all participating systems for English gender and second best for the joint prediction of gender and age.

6 Conclusions

We trained a linear SVM model that can predict gender and age groups for users based on what they write, and submitted this model to the Author Profiling shared task of PAN 2016. The model was trained on Twitter data, but tested on different genres, namely blogs, social media, and reviews. When developing the system, we were aware that training and test data would come from different sources, but we had no information regarding the actual nature of the test data. For this reason, we focused on selecting and engineering features that we thought would adequately characterise social media users independently on the genre-specific aspects of Twitter.

The results we obtained on the test data are substantially lower than those obtained in previous editions of author profiling at PAN. Although they are also lower than those observed on the cross-validated training set (likely due to our choice of ignoring Twitter-specific features), they are in fact state-of-the-art for this cross-genre task. While it will

[3] The organisers of PAN 2014 believe that the highest accuracies on Twitter might be due to the larger number of documents [8], an observation that in our case might relate to our lower performance on Dutch data.

be worth investigating other features and possibly other approaches to learning and to data representation, it is probably the nature of this task that makes it intrinsically difficult. In developing future author/user profiling systems, we should bear this issue in mind when aiming to design flexible and portable systems which perform well across genres.

References

1. Álvarez-Carmona, M.A., López-Monroy, A.P., Montes-y Gómez, M., Villaseñor-Pineda, L., Jair-Escalante, H.: INAOE's participation at PAN'15: Author profiling task. In: Proceedings of CLEF (2015)
2. Bird, Steven, E.L., Klein, E.: Natural Language Processing with Python. O'Reilly Media Inc (2009)
3. Brants, T.: TnT: a statistical part-of-speech tagger. In: Proceedings of the sixth conference on Applied natural language processing. pp. 224–231. Association for Computational Linguistics (2000)
4. 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)
5. 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)
6. Perkins, J.: Python 3 Text Processing with NLTK 3 Cookbook. Packt Publishing Ltd (2014)
7. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at PAN 2015. In: CLEF (2015)
8. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the author profiling task at PAN 2014. In: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (eds.) Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. CEUR Workshop Proceedings, vol. 1180, pp. 898–927. CEUR-WS.org (2014)
9. Rangel, F., Rosso, P., Moshe Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at PAN 2013. In: CLEF Conference on Multilingual and Multimodal Information Access Evaluation. pp. 352–365. CELCT (2013)
10. 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)
11. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of age and gender on blogging, vol. SS-06-03, pp. 191–197 (2006)
12. Schnoebelen, T.: Do you smile with your nose? Stylistic variation in Twitter emoticons. University of Pennsylvania Working Papers in Linguistics 18(2), 117–125 (2012)
13. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. pp. 173–180. Association for Computational Linguistics (2003)