Who Is Hot and Who Is Not? Profiling Celebs on Twitter

Notebook for PAN at CLEF 2019

Matej Martinc[1,2], Blaž Škrlj[1,2], and Senja Pollak[1,3]

[1] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[2] Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia
[3] Usher Institute of Population Health Sciences, Medical School, University of Edinburgh, UK
matej.martinc@ijs.si, blaz.skrlj@ijs.si, senja.pollak@ijs.si

Abstract We describe the system developed for the Celebrity profiling shared task of PAN 2019, capable of determining the gender, birthyear, occupation and fame of celebrities given their tweets. Our approach is based on a Logistic regression classifier and simple n-gram features. The best performance is achieved on the task of gender prediction, while predicting fame and occupation are slightly harder for the system. The worst performance is unsurprisingly achieved on the task of predicting birthyear, the hardest classification problem with seventy unbalanced classes. The proposed system was 3rd in the global ranking of PAN 2019 Celebrity profiling shared task.

1 Introduction

Author profiling (AP) is a field that deals with learning about the demographics and psychological characteristics of a person based on the text she or he produced. The most common tasks from the field include gender, age and language variety prediction but due to a large quantity of content available from social networks, the number of tasks is growing rapidly.

Most AP research is centered around a series of scientific events and shared tasks on digital text forensics, most popular being the series of scientific events and shared tasks called PAN (Uncovering Plagiarism, Authorship, and Social Software Misuse)[4]. The first PAN event took place in 2011 and the first AP shared task was organized in 2013 [12]. One of the most commonly addressed tasks in PAN is the prediction of an author's gender, although previous shared tasks also included tasks such as age, language variety and personality prediction [11,13]. This year, due to the availability of a new celebrity corpus [18], the number of attributes to predict has increased, and the task includes gender, age, fame and occupation prediction.

[4] http://pan.webis.de/

This paper describes our approach to the Celebrity profiling shared task of PAN 2019 [19], which involves the construction of four classification models for four distinct profiling traits on the celebrity corpus.

The rest of the paper is structured as follows: in Section 2 the findings from the related work are presented. Section 3 describes the corpus and how it was preprocessed. In Section 4 we present the feature engineering and classification methodology, while Section 5 presents the results. After a short Discussion (Section 6), we conclude the paper and present ideas for future work in Section 7.

2 Related Work

The first and most popular task addressed in the field of AP was gender prediction. It became a mainstream research topic with the work by Koppel et al. [5], who conducted experiments on a subset of the British National Corpus and found that women have a more relational writing style and men have a more informational writing style. While deep learning approaches have been recently prevailing in many natural language processing and text mining tasks, the state-of-the-art research on gender classification mostly relies on extensive feature engineering and traditional classifiers. For example, the winners of the PAN 2017 competition [2] used a Support vector machine (SVM) based system with simple features (word unigrams, bigrams and character three- to five-grams). Second ranked team [6] used a Logistic regression classifier and a somewhat more complicated combination of word, character and part-of-speech (POS) n-grams, sentiment from emojis, and character flooding as features. In PAN 2016, the best gender classification performance was achieved by [8], who employed a Logistic regression classifer and used word unigrams, word bigrams and character tetragrams features.

PAN 2016 AP shared task also dealt with age classification. The winners in this task [17] used a linear SVM model and employed a variety of features: word, character and POS n-grams, capitalization (of words and sentences), punctuation (final and per sentence), word and text length, vocabulary richness, hapax legomena, emoticons and topic-related words. On the other hand, none of the previous PAN tasks included prediction of fame and occupation. While we are not aware of any study which dealt with the celebrity fame prediction, we acknowledge the research of [1], who among other classification tasks also dealt with the prediction of text author's occupation on Spanish tweets. They evaluated several classification approaches (bag of terms, second order attributes representation, convolutional neural network and an ensemble of n-grams at word and character level) and showed that the highest performance can be achieved with an ensemble of word and character n-grams.

3 Dataset Description and Preprocessing

The training set for the PAN 2019 Celebrity profiling shared task consists of English tweets from 33,836 celebrities and contains labels for fame, gender, occupation and birthyear (details of a dataset structure are presented in Table 1 and Figure 1). The number of tweets per author is not constant and all classes are inbalanced. The label

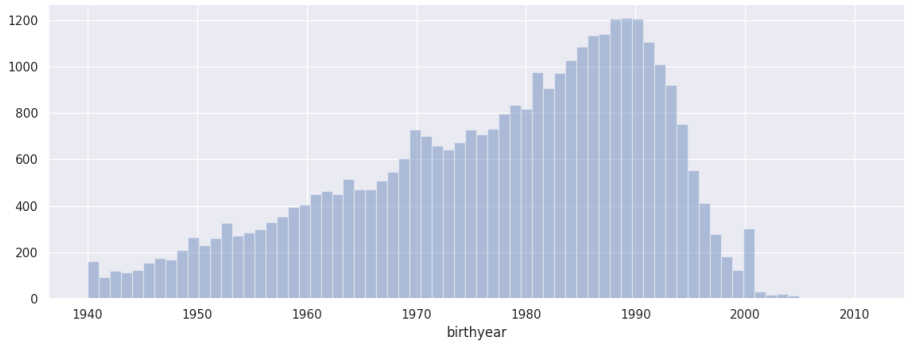

Figure 1. Birthyear distribution in the celebrity corpus

with the most classes is birthyear with 70 distinct values, occupation has 8 classes, while both fame and gender have 3 classes.

Table 1. Fame, gender and occupation distribution of the celebrity corpus

Fame	Gender	Occupation
superstar (7,116)	male (24,221)	sports (13,481)
star (25,230)	female (9.683)	performer (9,899)
rising (1,490)	non-binary (32)	creator (5,475)
/	/	politics (2,835)
/	/	science (818)
/	/	professional (525)
/	/	manager (768)
/	/	religious (35)

First, tweets belonging to the same celebrity are concatenated and used as one document in further processing. If an author has published more than 100 tweets, only first 100 tweets are used, since we believe this is a sufficient amount of content needed for successful profiling of the author and since this procedure drastically decreases the time and space complexity. After that, we employ three distinct preprocessing techniques on the resulting documents, producing three levels of preprocessed texts, which are all used in the feature engineering step:

- *Cleaned level*: replacing all hashtags, mentions and URLs with specific placeholders #HASHTAG, @MENTION, HTTPURL, respectively.
- *No punctuation level*: removing punctuation from the cleaned level;
- *No stopwords level*: stopwords are removed from the no punctuation level.

4 Feature Construction and Classification Model

Due to findings from the related work (see Section 2), which suggest that reliance on n-gram features and traditional classifiers is still the best approach for most author profiling tasks, we opted for the simplification of the approach we used in the PAN 2017 AP shared task [6]. According to the winners of the PAN 2017 competition [2], adding too sophisticated features negatively affects the performance of the author profiling classification model, therefore our model only contains three different types of n-gram features, which were normalized with the MinMaxScaler from the Scikit-learn library [9]:

- *word unigrams*: calculated on lower-cased No stopwords level, TF-IDF weighting (parameters: minimum document frequency = 10, maximum document frequency = 80%);
- *word bound character tetragrams*: calculated on lower-cased Cleaned level, TF-IDF weighting (parameters: minimum document frequency = 4, maximum document frequency = 80%);
- *suffix character tetragrams* (the last four letters of every word that is at least four characters long [14]): calculated on lower-cased Cleaned level, TF-IDF weighting (parameters: minimum document frequency = 10%, maximum document frequency = 80%).

We tested several classifiers from Scikit-learn [9]:

- Linear SVM
- SVM with RBF kernel
- Logistic regression
- Random forest
- Gradient boosting

An extensive grid search was performed in order to find the best hyper-parameter configuration for all tested classifiers and the best performing classifier was a Logistic regression with C=1e2 and fit_intercept= False parameters, same as in [6]. The Scikit-learn FeatureUnion[5] class was used to define prior weights for different types of features we used. The weights were adjusted with the help of the following procedure already described in [6]:

1. Initialize all feature weights to 1.0.
2. Iterate the list of features. For every feature repeat adding or subtracting 0.1 to the weight until the accuracy on the validation set is improving. When the best weight is found, move to the next feature on the list.
3. Repeat step 2 until the accuracy cannot be improved anymore.

The weights in our final Logistic regression model were the following:

- word unigrams and word bound character tetragrams: 0.8
- suffix character tetragrams: 0.4

[5] http://scikit-learn.org/stable/modules/generated/sklearn.pipeline.FeatureUnion.html

Table 2. Results on the unofficial validation set in terms of cRank (column All) and F1-score (all other columns)

Fame	Gender	Occupation	Birthyear	All
0.7837	0.9017	0.7578	0.0649	0.2092

5 Results

For the unofficial evaluation of our approach, the dataset was randomly split into train (containing 30,000 celebrities) and validation (containing 3,836 celebrities) sets. A separate classification model was trained for each of the classes and we measured performance of the model for each of the classes in terms of weighted F1-score. A measure used for the overall evaluation was cRank, which is a harmonic mean of models performance on each class, or formally:

$$cRank = \frac{4}{\frac{1}{F1_{fame}} + \frac{1}{F1_{gender}} + \frac{1}{F1_{occupation}} + \frac{1}{F1_{birthyear}}},$$

No lenience interval (as is the case in the official PAN 2019 Celebrity profiling shared task evaluation) was used for the birthyear F1-score calculation, therefore prediction was considered incorrect if the exact birthyear was not predicted. Results of the experiments for selected, best performing setting described in Section 4 on the unofficial validation sets in terms of F1-score and cRank are presented in Table 2.

Best results were achieved for the gender prediction task (F1-score of 90.17%), while the hardest attribute to predict was birthyear with an F1-score of only 6.46%. This is not surprising, due to a hard problem of classifying into 70 distinct unbalanced classes. Fame classification appears to be slightly easier for the classifier (F1-score of 78.37%) than the occupation prediction (F1-score of 75.78%) even though the occupation label has eight classes and fame only three. The overall cRank score is low (0.2092) due to the bad performance of the classifier on the task of birthyear prediction.

On the two official test sets the results are very different then on our unofficial validation sets (see table 3). F1-scores for fame, gender and occupation are about 30 percentage points lower on both official test sets, which suggests some serious overfitting. On the other hand, birthyear results on the two official test sets are about 30 percentage points better, most likely due to lenience interval used in the birthyear F1-score calculation, which also positively affected the overall cRank score. All in all, we ranked 3$^{\text{rd}}$ in the official TIRA [10] evaluation.

Table 3. Results on the two official test sets in terms of cRank (column All) and F1-score (all other columns)

	Fame	Gender	Occupation	Birthyear	All
Test dataset 1	0.517	0.580	0.449	0.361	0.462
Test dataset 2	0.507	0.594	0.486	0.347	0.465

6 Discussion

As in last years deep learning is gaining in popularity and achieving state-of-the-art results in a large variety of tasks [16,20,3] and as the celebrity corpus size is relatively large (compared to the PAN 2017 AP datasets), we also considered the neural transfer learning approach BERT (Bidirectional Encoder Representations from Transformers), proposed by [4]. Since the sequence length is limited to 512 characters, we decided to split the text document presenting tweets of each celebrity into chunks equal or smaller than 512 characters and used these chunks as training examples for the classifier. In the prediction phase, the classifier predicted labels for all chunks and majority voting was used to determine the final labels for the entire document. The initial experiments for gender and fame prediction however showed that the BERT classifier is performing much worse (achieving F1-scores of 83.33% and 72.11% for gender and fame, respectively) than the presented Logistic regression classifier. Thus, based on our experiments, we consider that traditional feature engineering techniques are still a better choice for the author profiling on PAN datasets.

7 Conclusion and Future Work

In this paper we have presented our approach to the PAN 2019 Celebrity profiling task, which deals with the prediction of fame, gender, occupation and birthyear for more than 30,000 celebrities. First, we present findings from the related work which suggest that a traditional classification approach with extensive feature engineering presented in this paper is still the preferred approach in the field of AP. We have tested several feature combinations and classifiers and finally selected a Logistic regression classifier with word unigram and character tetragram features, a system very similar to the one we proposed for the PAN 2019 Author profiling task [7].

The Logistic regression classifier and its hyper-parameters were chosen with a grid search but are identical to the study we conducted for the gender classification and language variety shared task in PAN 2017 [6], despite the celebrity corpus being almost ten times bigger than the PAN 2017 author profiling datasets. Because of the large dataset size we also tested the neural transfer learning approach proposed by [4], BERT (Bidirectional Encoder Representations from Transformers). The results were however worse than when the presented Logistic regression classifier was used. Our final results on the two official test sets are F1-scores of 51.7% for fame, 58.0% for gender, 44.9% for occupation and 36.1% for birthyear prediction on the first test dataset, and F1-scores of 50.7% for fame, 59.4% for gender, 48.6% for occupation and 34.7% for birthyear prediction on the second test dataset.

For future work, we believe investigation of potential semantic knowledge's effect on learning, such as explored in [21,15], could also provide valuable insights into parts of the feature space, relevant for learning. We also plan to evaluate the trained gender classification model on other AP datasets with gender labels which do not contain celebrities, in order to determine if the model is transferable.

Acknowledgments

The authors acknowledge the financial support from the Slovenian Research Agency core research programme Knowledge Technologies (P2-0103). The work of the second author was funded by the Slovenian Research Agency through a young researcher grant. This paper is also supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153 - project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

References

1. Aragón, M.E., López-Monroy, A.P.: Author profiling and aggressiveness detection in spanish tweets: Mex-a3t 2018. In: In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), CEUR WS Proceedings (2018)
2. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M.: N-gram: New groningen author-profiling model. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017. (2017)
3. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364 (2017)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), 401–412 (2002)
6. Martinc, M., Škrjanec, I., Zupan, K., Pollak, S.: Pan 2017: Author profiling-gender and language variety prediction. CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers (2017)
7. Martinc, M., Škrlj, B., Pollak, S.: Fake or not: Distinguishing between bots, males and females. Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019) (2019)
8. Modaresi, P., Liebeck, M., Conrad, S.: Exploring the effects of cross-genre machine learning for author profiling in PAN 2016. CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers (2016)
9. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research 12, 2825–2830 (2011)
10. Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA Integrated Research Architecture. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF. Springer (2019)
11. Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at PAN 2015. In: CLEF 2015 Working Notes. CEUR (2015)
12. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at pan 2013. Notebook Papers of CLEF pp. 23–26 (2013)
13. Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2017)

14. Sapkota, U., Bethard, S., Montes-y-Gómez, M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. In: NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015. pp. 93–102 (2015), http://aclweb.org/anthology/N/N15/N15-1010.pdf
15. Škrlj, B., Martinc, M., Kralj, J., Lavrač, N., Pollak, S.: tax2vec: Constructing interpretable features from taxonomies for short text classification. arXiv preprint arXiv:1902.00438 (2019)
16. Tang, D., Qin, B., Liu, T.: Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 1422–1432 (2015)
17. Busger op Vollenbroek, M., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., Haagsma, H., Nissim, M.: Gronup: Groningen user profiling notebook for PAN at clef 2016. CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers (2016)
18. Wiegmann, M., Stein, B., Potthast, M.: Celebrity Profiling. In: Proceedings of ACL 2019 (to appear) (2019)
19. Wiegmann, M., Stein, B., Potthast, M.: Overview of the Celebrity Profiling Task at PAN 2019. In: Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2019)
20. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1480–1489 (2016)
21. Škrlj, B., Kralj, J., Lavrač, N., Pollak, S.: Towards robust text classification with semantics-aware recurrent neural architecture. Machine Learning and Knowledge Extraction 1(2), 575–589 (2019), http://www.mdpi.com/2504-4990/1/2/34