Know your Neighbors: Efficient Author Profiling via Follower Tweets

Notebook for PAN at CLEF 2020

Boško Koloski[1,2], Senja Pollak[1], and Blaž Škrlj[1]

[1]Jožef Stefan Institute, Ljubljana
[2]Faculty of Information Science - University of Ljubljana
blaz.skrlj@ijs.si

Abstract User profiling based on social media data is becoming an increasingly relevant task with applications in advertising, forensics, literary studies and sociolinguistic research. Even though profiling of users based on their textual data is possible, social media such as Twitter offer also insight into the data of a given user's followers. The purpose of this work was to explore how such follower data can be used for profiling a given user, what are its limitations and whether performances, similar to the ones observed when considering a given user's data directly can be achieved. In this work we present our approach, capable of extracting various feature types and, via sparse matrix factorization, learn a dense, low-dimensional representations of individual persons solely from their followers' tweet streams. The proposed approach scored second in the PAN 2020 Celebrity profiling shared task, and is computationally non-demanding.

1 Introduction

User profiling on social media is becoming an increasingly relevant task when detecting problematic users or bots. In the era of social media, text-based representations of such users need to be learned, which is becoming a lively research area [5]. Online social media, such as Twitter, offer an unique opportunity to test to what extent properties of users can be predicted, and what potential implications of such learning endeavours are [3]. This paper discusses the challenge of predicting a given user's property based *solely* on the information captured from a given *user's followers' texts*. The paper explores to what extent the follower data offers profiling capabilities and what are its limitations. The schematic overview of the scenario considered in this work is shown in Figure 1. The remainder of this work is structured as follows. In Section 2 we present the related work, followed by the description of the proposed system (Section 4), experimental evaluation (Section 6) and the concluding remarks in Section 8.

2 Related Work

One of the first author profiling tasks was gender prediction by Koppel et al. [4], 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

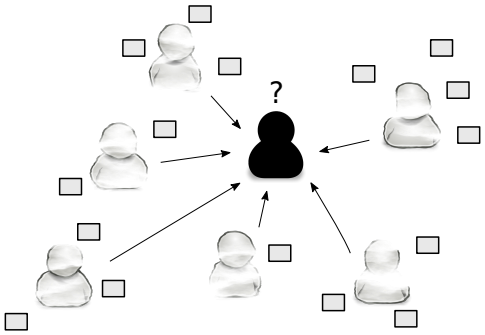


Figure 1: Schematic overview of the considered task. The gray boxes around the users surrounding the user of interest (middle node), are the users' tweets.

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.

Examples of previous PAN competition winners include [2] (who used support vector machines), however, the second ranked solution [7] was even simpler, employing only logistic regression classifier with features containing also emoji information and similar. 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 four-gram features.

PAN 2016 AP shared task also dealt with age classification. The winners in this task [12] used a linear SVM model and employed a variety of features: word, character and POS tag n-grams, capitalization (of words and sentences), punctuation (final and per sentence), word and text length, vocabulary richness, emoticons and topic-related words. We acknowledge also 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. Finally, the modeling task addressed in this work is similar to the last year's PAN Celebrity Profiling Challenge that aimed at predicting age, gender, fame and occupation[13], from which we also sourced some of the ideas used in the final models. The winning approach last year used tf-idf features with logistic regression and SVM classifiers [10].

3 Dataset Description and Preprocessing

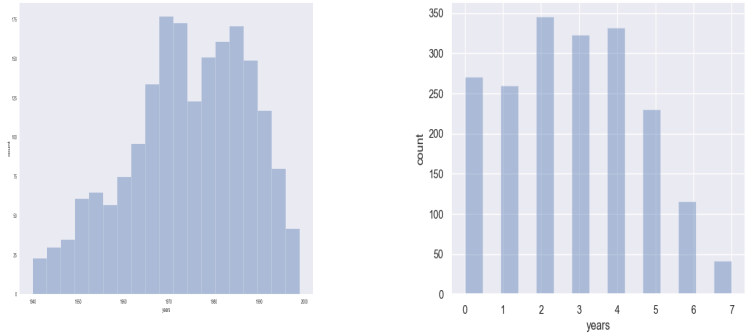
The training set for the PAN 2020 Celebrity Profiling shared task is composed of English tweets of follower feeds of 1,920 celebrities, labeled in three categories: gender,

occupation and birthyear. The dataset is balanced towards gender and occupation, while the birthyear label is not balanced. Distribution of the gender and occupation data is shown in Table 1 and birthyear data is presented in Figure 2 containing the original distribution and the augmented one, as described in Section 6.

Table 1: Distribution of gender and occupation labels.

Gender	Count
male	1072
female	1848

Occupation	Count
sports	480
performer	480
creator	480
politics	480



(a) Birthyears from the initial dataset

(b) Birthyears after augmenting the dataset

Figure 2: Overview of the birthyear distributions.

For getting the data prepared we firstly select 20 tweets for 10 authors for each celebrity, meaning 200 tweets in total for each celebrity in our data. Next, the tweet data is concatenated and preprocessed, as discussed next.

4 Feature Construction and Classification Model

The following section includes description of the proposed method and its intermediary steps.

Before feature construction, dimensionality reduction and classifier application, in the initial step we construct multiple representations of a given user that we denote as a collection C. The space of constructed features, similarly to [6] and [7], is based on:

– original text
– punctuation free - from the original text we removed punctuation
– stop-words free - from the punctuation free version stop words are removed

5 Authomatic feature seleciton

The collection C consists of multiple representations for each author, offering large space of potential features. We focused on character and word-level features to capture potentially interesting semantics. For this step, we used the SciKit-learn's [9] word tokenizer. The generated features are described as follows:

– character based - from each part in the collection C we generate character n-grams (up to 1 or 2 characters) and up to $\frac{n}{2}$ maximum allowed character features.
– word based- from each part in the collection C we generate word n-grams (up to 1,2,3 words) and up to $\frac{n}{2}$ maximum allowed word n-gram features

At the conclusion of the pipeline execution, we have prepared word and character features from each celebrity's collection of tweets, ready to be used in the feature selection step, which are finally joined via SciKit-learn's FeatureUnion.

5.1 Dimensionality reduction via matrix factorization

Finally, we perform sparse singular value decomposition (SVD)[1] that can be summarized via the following expression:

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T.$$

The final representation (embedding) $\boldsymbol{E}$ is obtained by multiplying back only a portion of the diagonal matrix ($\boldsymbol{\Sigma}$) and $\boldsymbol{U}$, giving a low-dimensional, compact representation of the initial high dimensional matrix. Note that $\boldsymbol{E} \in \mathbb{R}^{|D| \times d}$, where d is the number of diagonal entries considered. The obtained $\boldsymbol{E}$ is suitable for a given down-stream learning task, such as classification (considered in this work). Note that performing SVD in the text mining domain is also commonly associated with the notion of *latent semantic analysis*.

5.2 Classifier selection

For each sub task we performed extensive grid-search using [9] GridSearchCV and found classifiers that suited task the most. Following this goal we conducted a series of experiments, consisting of trying different environments and linear models as presented in the Section 6. Among the one we used were (SciKit learn's [9]) Support Vector Machines, Random Forests and Logistic Regression.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

6 Experiments

Series of experiments were executed in order to find the best embedding space and model. We explored various ways of modeling the birthyear variable:

- **R** - regression - where we applied linear regression and XGBoost Regressor [9] learner to derive a simple model to predict the years, where we predicted birthyear in the interval:

$$\max(1949, \min(\text{predicted\_year}, 1999)).$$

- **FC** - full classification - we applied classification learner to the task discrimination between each of the 60 classes (one year = one class)
- **AC** - altered classification - we applied classification to an altered label space where we reduced the number of labels to more balanced intervals, finally obtaining 8 of them, hence: 1949 - 1958, 1959 - 1966, 1967 - 1973, 1974 - 1980, 1981 - 1986, 1987 - 1991, 1992 - 1995, 1996 - 1999. For the final reverse prediction in the interval back we used the following estimates.
  1. predicting the middle of the interval
  2. predicting random year from the interval

For all tasks we considered GridSearchCV over parameter space to find best hyperparameter configuration, dimension number k and the number of features to be generated n. By doing 10-fold cross validation, the grid consisted of reducing the dimensions parametrized by k in the following interval:

$$k \in [128, 256, 512, 640, 768, 1024, 2048]$$

and the number of generated n features from the interval

$$n \in [2500, 5000, 10000, 20000, 30000, 50000].$$

The initial dataset was split to training(90%) and evaluation(10%) sets from after which we obtain C_{training} and $C_{\text{evaluation}}$. Once constructed, the feature space was considered for learning. We experimented with XGBoost, logistic regression and linear SVMs, of which hyperparameters we optimized in 5 fold cross validation. Finally, we tested the performance on the $C_{\text{evaluation}}$ set.

7 Results

This section includes the results of the empirical evaluation, used to select the final model. The obtained results are shown in table 2.

Table 2: Final evaluation on training data on TIRA.

name	#features	#dimensions	f1 age	f1 gender	f1 occupation	crank
model-**AC**-2	20000	512	**0.358**	**0.665**	0.656	**0.516**
model-**AC**-1	20000	512	0.346	0.663	**0.669**	0.509
model-**FC**-2	10000	512	0.313	0.639	0.632	0.473
model-**FC**-1	10000	512	0.291	0.605	0.648	0.452
model-**R**	10000	512	0.298	0.612	0.613	0.453
baseline-ngrams	#	#	0.362	0.584	0.521	0.469

The best scoring model is model-**AC**-2, which we chose for (final) test evaluation. Its hyperparameters were: $n = 20000$ features reduced to $k = 512$, while the Logistic Regression (occupation and age)'s regularization was set to $\lambda_2 = 1$. For gender, the SVM's hyperparameters were $\lambda_2 = 1$, gamma factor = scale and the polynomial kernel was used.

The best preforming model of experiments conducted in Section 6 yielded the following results on the test set on the TIRA site. We next present the official ranking of the proposed solution on the final TIRA test set.

TEAM	TEST-DATASET			
	CRANK	AGE	GENDER	OCCUPATION
baseline-ngram-celebrity-tweets	0.631	0.500	0.753	0.700
hodge20	0.577	0.432	0.681	0.707
koloski20	0.521	0.407	0.616	0.597
tuksa20	0.477	0.315	0.696	0.598
baseline-ngram-follower-tweets	0.469	0.362	0.584	0.521
random	0.333	0.333	0.500	0.250

Figure 3: The proposed submission achieved 2nd place (koloski20) (not accounting for full-tweet baseline).

The proposed system scored the second highest (the first listed in Figure 3 is the baseline based solely on a given author's tweet stream. It outperforms the generic baselines, whilst maintaining a lower dimension of the representation.

8 Discussion and conclusions

As not a single competing submission (Figure 3) achieved performance above the baseline trained on a given person's tweets, this task demonstrates that such type of classification is exceptionally hard, and needs to be fundamentally re-thought to overcome the full-information models aware of a given person's tweets. Significant improvement was achieved from the thresholding of the years and reducing the number of age classes to less than initial given, since the f1-score of age was based on the hit interval for years, giving us an uphold for varying different interval pooling strategies, namely we used two: first one based on generating the middle year in our predefined year interval and

the second was guessing a random number from the interval. The celebrity's own tweets and tweets of its followers gave competitive f1-scores while using relatively simple features (no emojis or similar) and computationally efficient methods representation construction methods. Finally, the score was calculated by calculating the harmonic mean of f1-scores:

$$cRank = \frac{1}{\frac{1}{f1\_occupation} + \frac{1}{f1\_birthyear} + \frac{1}{f1\_gender}}$$

As seen in the 7 section we believe that improving one the score on one subtask will only benefit the whole model if we keep or improve the scores on the other subtasks.

Further work will include trying out different division of the birthyear values by trying out different thresholds, possibly trying to inject more semantically enriched vectorization features [11] of tweets or improve the way the data is polled to build the data representation for a single celebrity.

9 Acknowledgements

The work of the last author was funded by the Slovenian Research Agency through a young researcher grant. The work was also supported by the Slovenian Research Agency (ARRS) core research programme *Knowledge Technologies* (P2-0103), an ARRS funded research project *Semantic Data Mining for Linked Open Data* (financed under the ERC Complementary Scheme, N2-0078) and EU 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. Batool, R., Khattak, A.M., Maqbool, J., Lee, S.: Precise tweet classification and sentiment analysis. In: 2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS). pp. 461–466. IEEE (2013)
4. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), 401–412 (2002)
5. Markov, I., Gómez-Adorno, H., Posadas-Durán, J.P., Sidorov, G., Gelbukh, A.: Author profiling with doc2vec neural network-based document embeddings. In: Mexican International Conference on Artificial Intelligence. pp. 117–131. Springer (2016)
6. Martinc, M., Blaž Škrlj Pollak, S.: Fake or not: Distinguishing between bots, males and. CLEF 2019 Evaluation Labs and Workshop – Working Notes Papers (2019)
7. 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)

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., 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)
10. Radivchev, V., Nikolov, A., Lambova, A.: Celebrity Profiling using TF-IDF, Logistic Regression, and SVM—Notebook for PAN at CLEF 2019. In: Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2019), http://ceur-ws.org/Vol-2380/
11. 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)
12. 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)
13. Wiegmann, M., Stein, B., Potthast, M.: Celebrity profiling. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 2611–2618 (2019)