Author Profiling: Bot and Gender Prediction using a Multi-Aspect Ensemble Approach

Notebook for PAN at CLEF 2019

Hamed Babaei Giglou[1] , Mostafa Rahgouy[1]*, Taher Rahgooy[2], Mohammad Karami Sheykhlan[1] ,and Erfan Mohammadzadeh [1]

[1] University of Mohaghegh Ardabili, Computer Science Department, Ardabil, Iran
* mostafarahgouy@student.uma.ac.ir
{hamedbabaeigiglou, mohammadkaramisheykhlan,
er.mohammadzadeh}@gmail.com
[2] Tulane University , Computer Science Department, New Orleans, LA, USA
trahgooy@tulane.edu

Abstract Author Profiling is one of the most important tasks in authorship analysis. In PAN 2019 shared tasks, the gender identification of the author is the main focus. Compared to the previous year the author profiling task is expended by having documents written by bots. In order to tackle this new challenge we propose a two phase approach. In the first phase we exploit the TF-IDF features of the documents to train a model that learns to detect documents generated by bots. Next, we train three models on character-level and word-level representations of the documents and aggregate their results using majority voting. Finally, we empirically show the effectiveness of our proposed approach on the PAN 2019 development dataset for author profiling.

Keywords: Author Profiling, User Modeling, Natural Language Processing, Supervised Machine Learning, Stacking ensemble.

1 Introduction

As computational power grows and artificial intelligence techniques evolve, new challenges, that were out of reach of machine learning in few years ago, emerge. One such a problem is author profiling. Different from the traditional authorship identification, in which a closed set of possible authors is known, author profiling aims to determine what are the characteristics of the authors: their age, gender, native language among others[13].

In the past years, multiple shared tasks have been organized on the topic of author profiling[11,8]. In this paper,we describe our approach for the Author Profiling shared task at PAN 2019 [9]. This year's Author Profiling task, is the 7th iteration of this task and is different from the previous years, since the data now includes bots.

The rest of the paper is organized as follows. Section 2 provides background and presents some related works on author profiling in general. Section 3 introduce our approach. Results are covered in Section 4. In Section 5 we draw a conclusion.

2 Related Work

The author attribution task is presented for the first time in PAN 2013 [10] in a single domain setting and continued to be part of PAN afterwards. AP task has been expanded and modified to more complicated and challenging tasks by considering multi-language settings, bot setting [9]. It has been shown that TF-IDF weighted n-gram features, both in terms of characters and words, are very successful in capturing especially gender distinctions[11].

In a similar work to our work, author of [3] undertook gender classification of Twitter users and achieved 76% accuracy, when trained their model only on tweets, using word unigram and bigrams and character 1- to 5-grams as features. In another related work [4] make use of LSA on the TF-IDF matrix with a linear SVM classifier. SVM has long proven a strong performance; however, they claimed that logistic regression is worth considering in similar experiments. In another related work [1] make use of traditional character n-gram analysis in combination with a linear SVM as a classifier. They found the optimal values for their model with dynamic and ad-hoc grid search approach and achieved reasonable results at PAN 2017.

3 Proposed Approach

In this section, we first describe how we pre-process the data, then we present the features we extract from the text. Finally, we describe the supervised models we use for classification.

3.1 Preprocessing

The first step in the proposed algorithm is to clean the input documents. In this step, we combine each user tweets and then removed the *Emojis*, *URLs*, *Mentions*, *Hashtags*, *Smileys*, *Numbers*, and *Reserved words(RT, FAV)* using *tweet-preprocessor*[1] python library. Next, using NLTK 3.0 Toolkit[2] we removed the stop-words from tweets and then lemmatize the words using WordNetLemmatizer(from NLTK).

[1] https://pypi.org/project/tweet-preprocessor/

3.2 Type Modeling

Term Frequency-Inverse Document Frequency (TF-IDF) was a common feature transformation technique for detecting author's style. First, we build a vocabulary using preprocessed train-set for each language with frequency term **5**. Next, using the scikit-learn [7] *TfidfVectorizer* method we convert a collection of raw tweets to a matrix of *TF-IDF* features. In final we used It TF-IDF in combination with a supervised method called a *linear SVM*. Based on work [1] which they are used the grid search tool shipped with scikit-learn python machine learning library to optimized values are listed in Table-1 so we used that optimized values to optimizing own type detection system. Since we didn't get much time to try different set of these parameters with different classifiers we only use these parameters with a linear SVM classifier. All runs were performed with *5*-fold cross-validation, while optimizing the *accuracy* target. Also, we used the Sub-linear term frequency scaling. which uses *1 + log(TF)* instead of TF.

Module	Parameters	Possible values
Feature Extraction	decode_error	**replace**
	norm	l1, **l2**
	script accents	true , **false**
	sublinear_tf	**True**, False
Transformation	Scaling	**MaxAbsScaler**
Classifier	C	**0.01** ,0.1 ,1 ,5
	kernel	**linear**

Table 1. Parameters obtained with grid search. Parameters with results in bold are optimal values.

3.3 Gender Modeling

Because the first part of the task is to identify the type of users and in case of human we have to identify the gender of users. If in the first part of the task we predict authors wrongly as a bot so we have to be precise to build an accurate system to compensation the wrong predictions of the first part of the task. Since ensemble methods usually produce more accurate solutions than a single model we used the ensemble method for gender profiling which uses majority voting for making a final prediction for the test instances.

Since we have two class here(male and female) so we use three different models for the ensemble method, *TF-IDF*[14], *N-gram*, and *Doc2Vec*[5].

TF-IDF for TF-IDF model we use the same way that we used for type modeling in the previous section with a linear SVM classifier.

N-gram To consider character level features we used character 4-gram frequencies in combination with a supervised method called a linear SVM method. The parameters which we used for 4-gram are in table 2. These are obtained by grid search tools from scikit-learn library. All runs were performed with 5-fold cross-validation, while optimizing the accuracy target.

Module	Parameters	Possible values
Feature Extraction	minimal document frequency	4, **5**
	n-gram order	3, **4**, 5
	lowercase	True , **False**
	script accents	True , **False**
Transformation	Scaling	**MaxAbsScaler**
Classifier	C	**0.01** ,0.1 ,1 ,5
	kernel	**linear**

Table 2. Parameters obtained with grid search. Parameters with results in bold are optimal values.

Doc2Vec Document embeddings is a Paragraph Vector, an unsupervised algorithm that learns fixed-length feature representations from variable-length pieces of texts, such as sentences, paragraphs, and documents. With regard to this idea and based on work [6] which they show that under certain settings the neural network-based features outperform the traditional features when using a *logistic regression* classifier. We train Doc2Vec for each language on train-set using Gensim[12] library and *MasAbsScaler* as scaling transformation for third model of ensemble method for gender identification.

3.4 Other Features

In the process of feature engineering, we explored many other ideas, some of which preformed poorly and thus we did not get them involved to our final approach. Yet, we feel some of them are worth mentioning.

Emoji Regard by our hypothesize emojis play an important role in identifying gender so ignoring emojis may lead to loss of valuable information about the author's gender. As a result, we used emoji in varies ways. Namely, we calculated the number of occurrences of emoji in a given document. In the other case, we used emoji as sentiment analysis features to distinguishing between male and female. However, using emoji as a feature caused to collapse our results thus we decided to ignore them.

Contracted Wordforms We use is based on the discrepancies in spelling for words that allow contracted forms, e.g.,I will (I'LL), are not (aren't). People typically favor one of the alternatives, and thus we use forms based on contracted apostrophes as discriminative features for detecting the gender of each author.

Quotation Marks Some authors may prefer either single or double quotation marks. We use the difference between the number of single and double quotes in a given document to see which group(male, female) prefer to use single or double quotation.

Sentence Length we noticed that some authors prefer to write long sentences where they use more conjunction in their text whereas some of them use short sentences as a result we consider length as a feature also we calculate the number of conjunctions in a given document and treat them as a feature.

Negations Another feature we use is based on the negation form e.g., impossible(not possible) to identifying the similarity of authors. We calculated the number of negations for each given document and added them to other features but we didn't get any notable result.

3.5 Final System

We submitted our final run, a linear SVM system for type identification(human and bots) which uses TF-IDF and in case of human we use ensemble system for gender identification(male and female) which uses TF-IDF , 4-gram with a linear SVM system, and Doc2Vec with a LogisticRegression system. This process visualized in the figure 1.

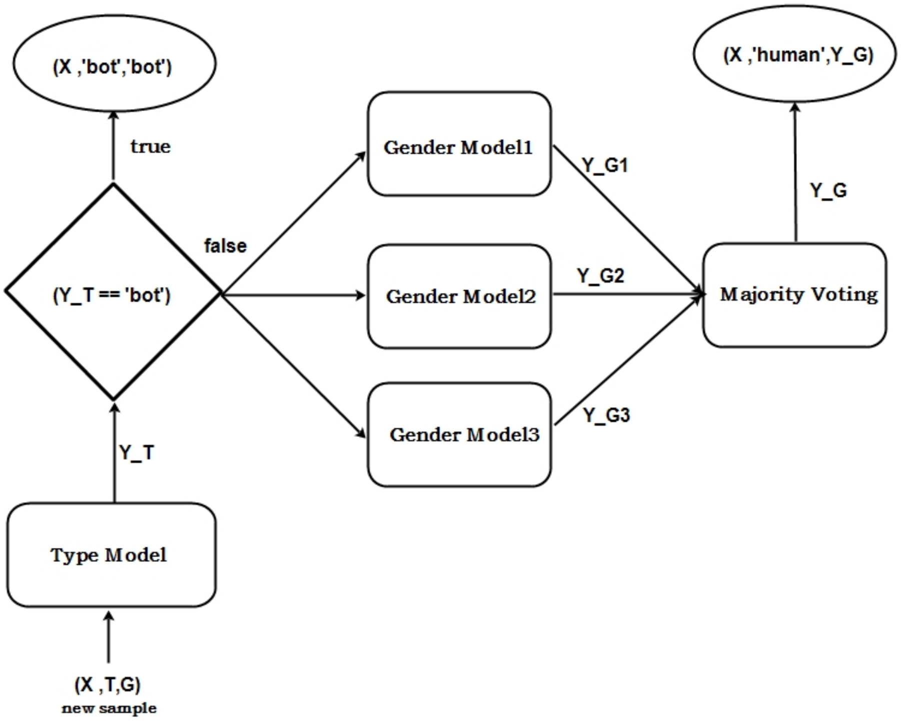

Figure 1. The architecture of the proposed approach.

4 Experimental Results

In Table 3, the performance of three models for gender identification wich described in section 3 are presented. It is clear that the ensemble of all three models is the best model in average.

Language	Gender Models			
	Models			Ensemble
	TF-IDF	N-gram	Doc2vec	TF-IDF + N-gram + Doc2Vec
EN	0.7906	0.7887	0.7709	**0.8032**
SP	**0.6543**	0.6347	0.5695	0.6521

Table 3. Accuracy results on development dataset for each different models for Gender Modeling.

Furthermore, we present the detailed results of the author profiling on development dataset of PAN 2019 in Table 4. As you can see, the average score obtained by this approach for English is **0.8330** and for Spanish is **0.7358** .

Language	Models						Average
	Type Model			Gender Model			
	TF-IDF	TP	Test Size	TF-IDF+N-gram+Doc2Vec	TP	Test Size	
EN	0.8629	1070	1240	0.8032	498	620	**0.8330**
SP	0.8195	750	920	0.6521	300	460	**0.7358**

Table 4. Detailed accuracy results on development dataset for each models.

5 Conclusion

In this paper, we proposed an algorithm for Author Profiling task in PAN 2019. We present our supervised approach, which uses a TF-IDF representation of the documents to distinguish between human and bots. Whereas, for detecting gender we present our ensemble approach which use a TF-IDF, N-gram, and Doc2Vec as feature extractions which makes predictions using an ensemble of diverse models including *LinearSVC* and *Logistic Regression*. Next, we uses dynamically determined parameters from an ad-hoc grid search approach to find the optimal values. Our evaluation shows that our approach is very capable of distinguishing between human and bot. However, identifying gender needs further work. The proposed algorithm implemented in *Python* and published on *GitHub*[2].

[2] https://github.com/HamedBabaei/PAN2019_bots_gender_profiling

References

1. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M.: N-GrAM: New Groningen Author-profiling Model—Notebook for PAN at CLEF 2017. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland. CEUR-WS.org (Sep 2017), http://ceur-ws.org/Vol-1866/
2. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc." (2009)
3. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1301–1309. EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2145432.2145568
4. Daneshvar, S., Inkpen, D.: Gender Identification in Twitter using N-grams and LSA—Notebook for PAN at CLEF 2018. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) CLEF 2018 Evaluation Labs and Workshop – Working Notes Papers, 10-14 September, Avignon, France. CEUR-WS.org (Sep 2018), http://ceur-ws.org/Vol-2125/
5. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: ICML (2014)
6. Markov, I., Gómez-Adorno, H., Posadas-Durán, J.P., Sidorov, G., Gelbukh, A.F.: Author profiling with doc2vec neural network-based document embeddings. In: MICAI (2016)
7. 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)
8. Potthast, M., Pardo, F.M.R., Tschuggnall, M., Stamatatos, E., Rosso, P., Stein, B.: Overview of pan'17 - author identification, author profiling, and author obfuscation. In: CLEF (2017)
9. Rangel, F., Rosso, P.: Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling. In: Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2019)
10. Rangel Pardo, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the Author Profiling Task at PAN 2013. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain. CEUR-WS.org (Sep 2013)
11. Rangel Pardo, 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: Balog, K., Cappellato, L., Ferro, N., Macdonald, C. (eds.) CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal. CEUR-WS.org (Sep 2016), http://ceur-ws.org/Vol-1609/16090750.pdf
12. 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
13. Rodrigo Ribeiro Oliveira, R.d.: Using Character n-grams and Style Features for Gender and Language Variety Classification—Notebook for PAN at CLEF 2017. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland. CEUR-WS.org (Sep 2017), http://ceur-ws.org/Vol-1866/
14. Salton, G., McGill, M.J.: Introduction to modern information retrieval (1986)