# Author Profiling on Social Media: An Ensemble Learning Model using Various Features

## Notebook for PAN at CLEF 2019

Youngjun Joo and Inchon Hwang

Department of Computer Engineering
Yonsei University, Seoul, Korea
{chrisjoo12, ich0103}@gmail.com

**Abstract** We describe our participation in the PAN 2019 shared task on author profiling, determine whether a tweet's author is a bot or a human, and in case of human, identify author's gender for English and Spanish datasets. In this paper, we investigate the complementarities of both stylometry methods and content-based methods, putting forward various techniques for building flexible features. Acting as a complement to these methods, we investigate an ensemble learning method paves the way to improve the performance of AP tasks. Experimental results demonstrate that the ensemble method by the combination of the stylometry methods and content-based methods can more accurately capture the author profiles than traditional methods. Our proposed model obtained 0.9333 and 0.8352 of accuracy in the bot and gender identification tasks for English test dataset respectively.

## 1 Introduction

The Author profiling (AP) deals with the classification of shared content in order to predict general or demographic attributes of authors such as gender, age, personality, native language, or political orientation, among others [17]. Being able to infer an author's profile has wide applicability and has proved to be advantageous in many areas such as marketing, forensics, and security, etc.

Broadly speaking, the approaches that tackle AP view the task as a multi-class or single-label classification problem, when the set of the class label is known a priori [20]. Thus, AP is modeled as a classification task, in which automatic detection methods have to assign labels (e.g., male, female) to objects (texts). Consequently, most work has been devoted to determining a suitable set of features to deal with the task on the writing profile of authors. In the 2019 shared AP task on PAN dataset [2], the goal is to infer whether the author of a Twitter feed is a bot or a human and to profile the author's gender in case of human [16]. Both training and test data is provided in two different languages: English, Spanish.

In order to predict bot and gender, we propose the complementarities of both stylometry methods and content-based methods, putting forward various techniques for building flexible features (basic count features, psycholinguistic features, TF-IDF, Doc2vec). Acting as a complement to these features, we also investigate an ensemble learning method combining classification methods based on various features and BERT model paves the way to improve the performance of AP tasks.

## 2 Related Works

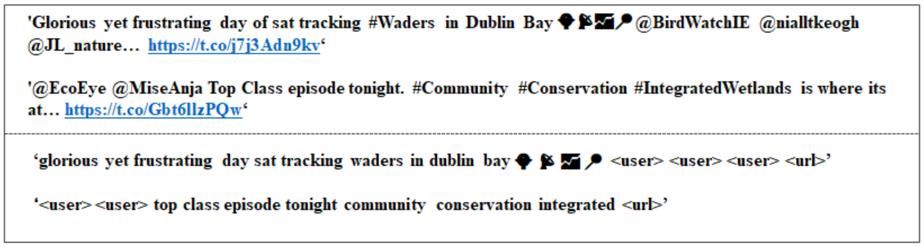
Approaches for predicting an AP can be broadly categorized into two types of methods: (1) stylometry methods which aim to capture an author's writing style using different statistical features (e.g., functional words, POS, punctuation marks, and emoticons), (2) content-based methods that intend to identify an author's profiles based on the content of the text (e.g., bag of words, words n-gram, term vectors, TF-IDF n-grams, slang words, emotional words), and topics discussed in the text (e.g., topic models such as LDA, PLSA ) [5]. According to the PAN[1] competitions, most successful works for AP in social media have used combinations of these two kinds of features.

Every author's writing style can be used to identify an author's attributes. In previous studies, style based features were used to predict the author's attributes, age, and gender [3,6,13,18,22]. In these methods, lexical word-based features represent text as a sequence of tokens forming sentences, paragraphs, and documents. A token can be the numeric number, alphabetic word or a punctuation mark. Plus, these tokens are used to statistics such as average sentence length, average word length, a total number of words and a total number of unique words, etc. Also, character-based features consider the text as a sequence of characters.

Content-based methods employ specific words or special content which are used more frequently in that domain than in other domains [23]. These words can be chosen by correlating the meaning of words with the domain [8, 23] or selecting from corpus by frequency or by other feature selection methods [1]. An analysis of information gain presented in [19] showed that the most relevant features for gender identification are those related with content words (e.g., *linux* and *office* for identifying males, whereas *love* and *shopping* for identifying females).

Recently, some works have used deep learning models and learning method of representations for AP [7, 9, 11, 21]. [7] used the approach based on subword character n-gram embeddings and deep averaging networks (DAN). [9] used the model consists of a bi-RNN implemented with a Gated Recurrent Unit (GRU) combined with an Attention mechanism. [11] proposed two models for gender identification and the language variety identification of four languages that consist of multiple layers to classify an author's profile trait with neural networks.

[1] https://pan.webis.de/



'Glorious yet frustrating day of sat tracking #Waders in Dublin Bay 🌳🐦📈📍 @BirdWatchIE @nialltkeogh @JL_nature… https://t.co/j7j3Adn9kv'

'@EcoEye @MiseAnja Top Class episode tonight. #Community #Conservation #IntegratedWetlands is where its at… https://t.co/Gbt6llzPQw'

'glorious yet frustrating day sat tracking waders in dublin bay 🌳 🐦 📈 📍 <user> <user> <user> <url>'

'<user> <user> top class episode tonight community conservation integrated <url>'

**Figure 1.** Illustration of text preprocessing and postprocessing.

# 3 Proposed Approach

## 3.1 Text Preprocessing and Postprocessing

The preprocessing of the text data is an essential step as it makes the raw text ready for applying machine learning algorithms to it. The objective of this step is to clean noise those are less relevant to detect the AP on the texts.

We, at first, aggregate tweet posts published by an individual user into one document before training to alleviate the shortcomings of short texts. In order to utilize most of the information in text, we perform some transforming tasks of the short texts (i.e., XML parsing, contradictions unfolding, text tokenizing, stemming, lemmatization, and removing stopwords). Also, for utilization and word-level representation on most of the text information, we perform spell correction for informal words using SymSpell library[1], word segmentation for splitting hashtags using WordSegment library[2], and annotation (surround or replace with special tags such as *<money>, <number>, <date>, <phone>*, or *<user>*) as a text postprocessing task (see Figure 3).

## 3.2 Basic Count Features

Previous works on AP tasks explore lexical, syntactic, and structural features. Lexical features are used to measure the habit of using characters and words in the text. The commonly used features in this kind consist of the number of characters, word, a frequency of each type of characters, etc. Syntactic features include the use of punctuations, part-of-speech (POS) tags, and functional words. Structural features represent how the author organizes their documents or other special structures such as greetings or signatures.

As shown Table 1, we construct a basic count feature set including punctuation characters (e.g., question marks, and exclamation marks) and other features (e.g., average syllable per word, functional word count, special character count, capital ratio, etc.).

[1] https://github.com/wolfgarbe/SymSpell

[2] https://github.com/grantjenks/python-wordsegment

**Table 1.** List of basic count features extracted from the Twitter.

| | |
|---|---|
| Average sentence lenght by character | Average sentence lenght by word |
| Average syllable per word | Average word length |
| Capital ratio | Number of functional words |
| Number of special characters | Character count |
| Word count | Hashtag count |
| Direct tweet count | Repeat punctuation count |
| Punctuation count | URL count |
| Positive emoji count | Negative emoji count |
| Emoji count | Slang words count |
| Stopword count | <money> tag count |
| <phone> tag count | <user> tag count |
| <time> tag count | <date> tag count |
| <number> tag count | |

### 3.3 Psycholinguistic Features

The relationship between personality traits and the use of language has been widely studied by the psycholinguist Pennebaker. [12] analyzed how the use of the language varies depending on personal traits. For example, in regards to the authors' gender, he found out that in English women use more negations or first persons, because they are more self-conscientious, whereas men use more prepositions in order to describe their environment. These findings are the basis of LIWC[1] (Linguistic Inquiry and Word Count) that is one of the most used tools for capturing people's social and psychological states, which have proved to be useful in the AP task.

LIWC has two types of categories; the first kind captures the writing style of the author like the POS frequency or the length of the used words (i.e., summary language variables, linguistic dimensions, other grammar). The second category (i.e., psychological processes) captures content information by counting the frequency of words related to some thematic categories such as affective processes, social processes, personal concerns, etc. Regarding the use of this tool, we focused on the content information.

### 3.4 TF-IDF

We adapted the TF-IDF (Term Frequency-Inverse Document Frequency) method to judge the topics of each document by the words it contains. In the TF-IDF words are given weight – TF-IDF measures relevance, not frequency. Particularly, word counts are replaced with TF-IDF scores across the corpus. TF-IDF, at first, measures the number of times that words appear in a given document (i.e., *term frequency*). However, since words such as *and* or *the* appear frequently in all documents, those must be systematically discounted (i.e., *inverse-document frequency*). The more documents a word appears in, the less valuable that word is as a signal to differentiate any given document. That is intended to leave only the frequent and distinctive words as markers. TF-IDF relevance of each word is a normalized data format as the following formula:

[1] http://liwc.wpengine.com/

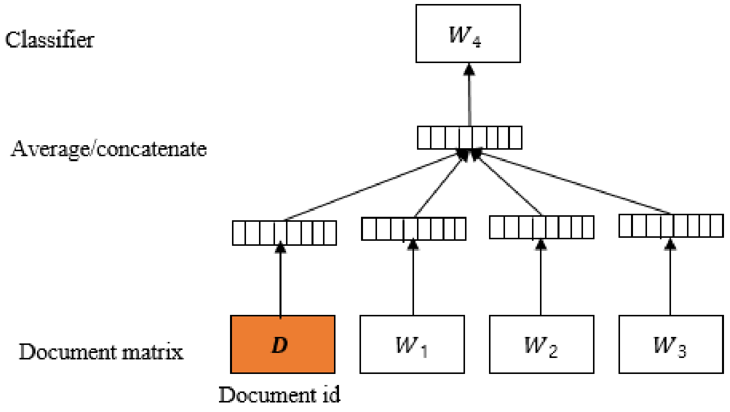


**Figure 2.** Framework for learning document vectors. Adpated from [10].

$$W_{i,j} = tf_{i,j} * log\,(\frac{N}{df_i}) \tag{1}$$

where $tf_{i,j}$ is the number of occurrences of $i$ in $j$ ; $df_i$ is the number of documenting containts $i$; $N$ is the total number of documents.

We extract unigrams, bigrams, and trigrams derived from the bag of words representation of each Twitter posts. To account for occasional differences in content length between train dataset and test dataset, these features are encoded as TF-IDF values.

### 3.5 Doc2vec

We use the Doc2vec method [10], an unsupervised learning model that learns feature representations of fixed length from the document of variable length. The idea is to combine the meaning of words for construction of the meaning of documents using a distributed memory model (see Figure 2). There are two models for the distributed representation of documents: Distributed Memory (DM) and Distributed Bag-of-Words (DBOW). The distributed representation obtained by this model outperforms both Bag-of-Words (BoW) and word n-gram models producing the new state of the art result for text classification and sentiment analysis tasks.

In this work, different representations of the texts were used as input data types for the Doc2vec method in order to evaluate the quality of different distributed representation outputs. In particular we represented the texts in terms of *word unigram*, *bigram* and *trigram*. For the implementation of the Doc2vec model, a freely available package of the Doc2vec included in the Gensim module used. The Doc2vec method offers two possible approaches (i.e., PV-DM, PV-DBOW) to build the model. As shown in Table 2, our experimental results report better performance when both representations are concatenated, thus our model's final document vector is composed of the concatenation of the representations obtained by the DM (Distributed Memory) and the DBOW (Distributed Bag of Words) models.

**Table 2.** Doc2vec results of gender classification on English train dataset.

| Doc2vec model | Dim | Acc |
|---|---|---|
| DBOW 1-gram | 300 | 0.8447 |
| DBOW 1+2-gram | 300 | 0.8500 |
| DBOW 1+2+3-gram | 300 | 0.8462 |
| DM 1-gram | 300 | 0.8402 |
| DM 1+2-gram | 300 | 0.8455 |
| DM 1+2+3-gram | 300 | 0.8438 |
| DBOW+DM 1-gram | 600 | 0.8606 |
| DBOW+DM 1+2-gram | 600 | **0.8639** |
| DBOW+DM 1+2+3-gram | 600 | 0.8618 |

### 3.6 BERT

BERT (Bidirectional Encoder Representations from Transformers) model's key innovation is applying the bidirectional training of Transformer, a popular attention model, to language modelling [4]. The Transformer includes two separate mechanisms – an encoder that reads the text input and a decoder that produces a prediction for the task. Since the goal of BERT is to generate a language model, only the encoder mechanism is necessary. As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left), the Transformer encoder reads the entire sequence of words at once. This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word).

When training language models, there is a challenge of defining a prediction goal. To overcome this challenge, BERT uses two training strategies: Masked LM (MLM) and the next sentence prediction (NSP). Their results show that BERT outperforms the state-of-the-art results in a wide variety of NLP tasks, including QA (SQuAD), Natural Language Inference (MNLI), and others. In this task, we used pre-trained BERT model (i.e., bert_uncased_L-24_H-1024_A-16) in TensorFlow Hub website [1].

## 4 Experimental Results and Discussion

The task of identifying bot and gender from the text is cast as a classification task. For these tasks, we have performed binary classification task, i.e., the goal is to distinguish between two classes: (1) bot and human (2) male and female in case of human class. We have used 10-fold cross validation for experiments. We employed Gradient Boosted Decision Tree (GBDT) classifier and BERT model to train and test our proposed system as classifier models. Our results on PAN2019 subtasks are summarized in Table 3. Regarding the submissions to the task through the TIRA [14] platform which is the web service platform to facilitate software submissions into the virtual machine, training was conducted offline by concatenating the training and validation sets as input and

[1] https://tfhub.dev/

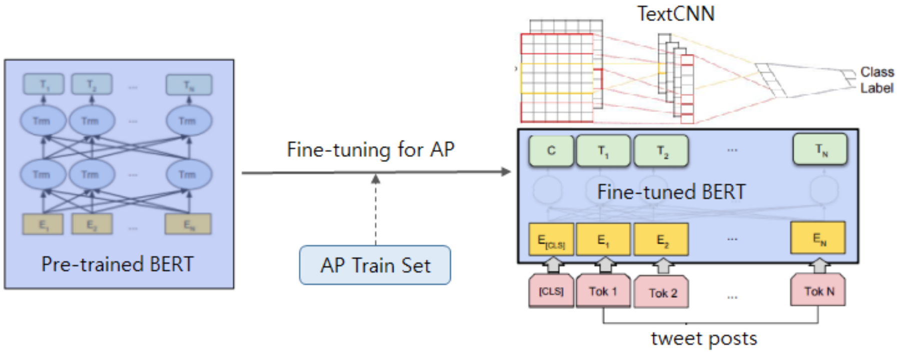


**Figure 3.** Architecture of BERT model for author profiling tasks.

**Table 3.** Experimental results for bot and gender classifications on PAN2019 dataset.

| No | Features | # | English | | Spanish | |
|---|---|---|---|---|---|---|
| | | | Bot | Gender | Bot | Gender |
| - | Baseline (LDSE [15]) | - | 0.9054 | 0.7800 | 0.8372 | 0.6900 |
| 1 | Basic count features | 25 | 0.9741 | 0.8152 | 0.9498 | 0.8373 |
| 2 | Psycholinguistic features | 56 | 0.9514 | 0.8623 | 0.8844 | 0.7310 |
| 3 | TF-IDF | 5000 | 0.9466 | 0.8379 | 0.9253 | 0.7657 |
| 4 | Doc2vec (1+2+3-gram) | 600 | 0.9498 | 0.8639 | 0.9132 | 0.7896 |
| 5 | Feature stacking(1+2+3+4) | 5681 | 0.9805 | 0.8701 | 0.9511 | 0.8547 |
| 6 | BERT (pre-trained) | - | 0.9902 | 0.8324 | - | - |
| 7 | Ensemble (5+6) | - | 0.9832 | 0.8801 | - | - |
| 9 | TIRA Test | - | **0.9333** | **0.8352** | N/A | N/A |

then, the trained models were deployed to TIRA to classify the unseen test dataset. We tuned our single models on the development datasets and submitted the final results using ensemble models. The ensemble models were built by averaging the outputs of two models, which are feature stacking model and deep learning model. Our submission results achieve an accuracy of **0.9333** for bot detection task and **0.8352** for gender identification task on English test dataset.

In the gender identification tasks, the content-based approaches show better results, especially the Doc2vec model shows significantly better results than other feature sets. The resulting accuracy scores by other features (i.e., *Basic count features, psycholinguistic features, TF-IDF*) indicate that these approaches address the problem to some extent but requires more distinctive features to further improve the accuracy. Interestingly, the concatenation of both stylometry and content-based approach proves highly effective, achieving the best results overall. However, Doc2vec method proves to be highly competitive as a single-handed feature set.

In terms of deep learning approach, even if BERT model managed to learn important features from basically any data structure without having to manually derive features, the performance in our experiment is not competitive. Our experimental results using the pre-trained model did not outperform other model's accuracy and seem to require more fine-tuning.

## 5 Conclusions

This paper summarized our participation in PAN2019 shared task on AP, where we aimed to deal with the challenge of AP. We tried to approach the problem from the perspective of stylometry and content-based methods, as well as contextualized word embeddings from BERT model. In terms of training method, we adopted an ensemble approach and carried on experiments on PAN collections. Experimental results demonstrate that the ensemble method by the combination of the style-based and content-based methods can more accurately capture the author profiles than traditional methods.

There is still room to improve our work in the future. In our system, we compared the BERT model showing state-of-the-art results in various NLP tasks without sophisticated fine-tunising. We will leave more fine-tuning work on our datasets as future work.

## References

1. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. Commun. ACM 52(2), 119–123 (2009)
2. Daelemans, W., Kestemont, M., Manjavancas, E., Potthast, M., Rangel, F., Rosso, P., Specht, G., Stamatatos, E., Stein, B., Tschuggnall, M., Wiegmann, M., Zangerle, E.: Overview of PAN 2019: Author Profiling, Celebrity Profiling, Cross-domain Authorship Attribution and Style Change Detection. In: Crestani, F., Braschler, M., Savoy, J., Rauber, A., Müller, H., Losada, D., Heinatz, G., Cappellato, L., Ferro, N. (eds.) Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019). Springer (Sep 2019)
3. De-Arteaga, M., Jimenez, S., Duenas, G., Mancera, S., Baquero, J.: Author profiling using corpus statistics, lexicons and stylistic features. Online Working Notes of the 10th PAN evaluation lab on uncovering plagiarism, authorship. and social misuse, CLEF (2013)
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. Fatima, M., Hasan, K., Anwar, S., Nawab, R.M.A.: Multilingual author profiling on facebook. Information Processing & Management 53(4), 886–904 (2017)
6. Flekova, L., Preoţiuc-Pietro, D., Ungar, L.: Exploring stylistic variation with age and income on twitter. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 313–319 (2016)
7. Franco-Salvador, M., Plotnikova, N., Pawar, N., Benajiba, Y.: Subword-based deep averaging networks for author profiling in social media. In: CLEF (Working Notes) (2017)
8. Hsieh, F., Dias, R., Paraboni, I.: Author profiling from facebook corpora. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018) (2018)
9. Kodiyan, D., Hardegger, F., Neuhaus, S., Cieliebak, M.: Author profiling with bidirectional rnns using attention with grus: Notebook for pan at clef 2017. In: CLEF 2017 Evaluation Labs and Workshop–Working Notes Papers, Dublin, Ireland, 11-14 September 2017. vol. 1866. RWTH Aachen (2017)

10. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International conference on machine learning. pp. 1188–1196 (2014)
11. Miura, Y., Taniguchi, T., Taniguchi, M., Ohkuma, T.: Author profiling with word+ character neural attention network. In: CLEF (Working Notes) (2017)
12. Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K.: The development and psychometric properties of liwc2015. Tech. rep. (2015)
13. Pervaz, I., Ameer, I., Sittar, A., Nawab, R.M.A.: Identification of author personality traits using stylistic features: Notebook for pan at clef 2015. In: CLEF (Working Notes) (2015)
14. 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)
15. Rangel, F., Franco-Salvador, M., Rosso, P.: A low dimensionality representation for language variety identification. In: The 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLingâĂŹ16). pp. 156–169. Springer-Verlag (2018)
16. 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)
17. Rangel, F., Rosso, P., Montes-y Gómez, M., Potthast, M., Stein, B.: Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. Working Notes Papers of the CLEF (2018)
18. Santosh, K., Bansal, R., Shekhar, M., Varma, V.: Author profiling: Predicting age and gender from blogs. Notebook for PAN at CLEF 2013 (2013)
19. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In: AAAI spring symposium: Computational approaches to analyzing weblogs. vol. 6, pp. 199–205 (2006)
20. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR) 34(1), 1–47 (2002)
21. Sierra, S., Montes-y Gómez, M., Solorio, T., González, F.A.: Convolutional neural networks for author profiling. Working Notes of the CLEF (2017)
22. Wanner, L., et al.: A semi-supervised approach for gender identification. In: Calzolari N, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S. LREC 2016, Tenth International Conference on Language Resources and Evaluation; 2016 23-28 May; Portorož, Slovenia.[Place unknown]: LREC, 2017. p. 1282-7. LREC (2016)
23. Zheng, R., Qin, Y., Huang, Z., Chen, H.: Authorship analysis in cybercrime investigation. In: International Conference on Intelligence and Security Informatics. pp. 59–73. Springer (2003)