Social Network Analysis and Mining (2021) 11:59
https://doi.org/10.1007/s13278-021-00768-6

ORIGINAL ARTICLE

Toward a new approach to author profiling based on the extraction of statistical features

Sarra Ouni[1] · Fethi Fkih[1,2] · Mohamed Nazih Omri[1]

Received: 31 December 2020 / Revised: 4 June 2021 / Accepted: 7 June 2021 / Published online: 22 June 2021

Abstract
Recently, author profiling on social media and on online platforms, characterized by a huge volumes of data, has become more than a critical issue. This issue is of increasing interest in various fields related to forensic medicine, security, marketing, education, etc. The main objective of author profiling is to identify the type of writer of the messages, whether it is a human or a bot with a very strong presence. These bots have the task of drawing the attention of browsers to specific events, often used to disseminate incorrect and/or false information. In this work, we offer a new approach to detect these bots and the kind of anonymous perpetrators on these social networks. Our approach, purely statistical, is based on digital features (APSF), extracted from users' tweets, and on the technique of random forests. A total of 17 stylometry-based features were used to train the model. To assess the performance of our approach, we considered different standard measures, namely accuracy, precision, recall and F1-score. The results obtained show that our approach gives the best performance for both English and Spanish languages. For the English dataset, we achieved an accuracy of 92.45% for the bot detection task and 90.36% for the gender classification; similarly, we obtained accuracy values of 89.68% and 88.88% for the Spanish dataset.

Keywords Bot detection · Gender detection · Features extraction · Machine learning · Twitter

1 Introduction

The rapid growth of social media networks such as Facebook, Twitter and Instagram has given the opportunity to users to publish content, communicate and exchange information with others in a very fast way. Often, social media users hide their real identity such as their true names, genders and also their occupations in order to express their opinions freely. Some other users mask their real information for fraudulent and other erroneous acts. The huge amounts of unstructured texts and the easiness of posting content in social media networks make the task of identifying the user's category (i.e., human, false profile, bot, etc.) difficult. Therefore, it has become very important to provide effective tools and methods to identify the true identity of users. An interesting field that has attracted the attention of several researchers is author profiling (AP). AP is a subtask of authorship analysis whose goal is to identify demographic attributes of authors such as age (Rangel et al. 2014), gender (Rangel et al. 2013) and personality traits (Pardo et al. 2015) by examining his/her written text. AP has a growing importance in several applications related to forensics purpose, security (Juola 2015), marketing (Pardo et al. 2015), psychology and terrorism prevention.

Twitter has become increasingly popular in the last two decades (Sendi et al. 2017). Twitter is a "microblogging" system that allows us to send and receive short and informal texts called tweets. It has a large number of users. In recent years, bots have become ubiquitous in Twitter (Chu et al. 2012; Davis et al. 2016; Subrahmanian et al. 2016). Twitter bots can be used for helpful purposes, such as broadcasting important content in different areas such as ideology, politics and marketing. They have been also used for different malicious purposes, such as spreading fake information, spamming, violating others' privacy and distributing

✉ Sarra Ouni
sarraouni93@gmail.com

Fethi Fkih
f.fki@qu.edu.sa

Mohamed Nazih Omri
mohamednazih.omri@eniso.u-sousse.tn

[1] MARS Research Lab LR 17ES05, University of Sousse, Sousse, Tunisia

[2] Department of Computer Science, College of Computer, Qassim University, Buraydah, Saudi Arabia

malware. In Bessi and Ferrara (2016), research shows that in 2016 US presidential Election, more than a one-fifth of tweets on Twitter came from bot accounts. Therefore, detection of social bots, especially Twitter bots, has become an important research domain across the globe for a variety of security purposes. As a result, it is important to develop AP systems capable of distinguishing bot profiles from human ones, as well as identify the gender of a human author. In this article, we present our approach for the PAN 2019 AP challenge (Pardo and Rosso 2019). The task here is to determine whether a tweet's author is a bot or a human and also to classify the gender of human users as male or female for English and Spanish datasets. Further details about our method are presented in the following sections.

1.1 Goal and contributions

The study that we propose in this article introduces the problem of AP in social networks, especially in Twitter. This work aims to distinguish bots from humans and in case of a human, to predict the gender of authors in Twitter based on statistical features. The goal of this work is to establish which stylometry-based features help to best address this challenge. For this purpose, we used the corpus from the AP task of PAN 2019 (Pardo and Rosso 2019). In this context, various well-known machine learning algorithms have been used. The random forest technique was the best performing for this work. The main contributions of this paper are outlined below:

- We propose a language-independent system for AP based on statistical features. Specifically, we provided empirical evidence regarding the pertinence of the style-based features for solving the posed task across Twitter;
- We evaluate the performance of our approach using machine learning techniques;
- We identify the most suitable classifier for AP application using statistical features;
- We discuss three state-of-the-art methods proposed to address the same task. Additionally, we compare their performance against the results of our proposed approach.

1.2 Paper structure

The remainder of this article is structured as follows. We start by giving a brief overview of existing work on bots detection problem and gender identification tasks in Sect. 2. Section 3 presents the motivations for the proposed method. In Sect. 4, we introduce our approach in which we explain how we preprocessed the data and what types of features could be considered as significant indicators for both gender and bot detection, we also present different classifiers used in our implemented model and we detail the proposed algorithm used for the classification tasks. Section 5 presents the experimental setup. We precisely present the used dataset, the strategy of defining the hyperparameters for all machine learning algorithms used. The evaluation metrics employed are also described in this part, and the obtained experimental results are discussed. Section 6 is reserved to present some limits of our approach. Section 7 concludes the paper and suggests some future directions.

2 Related work

By analyzing users posts on social networks, several researchers were able to determine different users' aspects such as gender, age and language. In recent years, the interest in the bot and gender profiling field of research has been increasing (Varol et al. 2017; Rangel et al. 2018), because numerous users tend to share false information, hate publications and fake news. In particular, bots disseminate often incorrect and false information. Therefore, their detection in social media is very much important. In Sect. 2.1, we provide a review of previous studies on the automatic bot detection task. In Sect. 2.2, other studies related to automatic gender classification task are presented. We mainly focus on the different features used in each study as well as the learning algorithms employed in the classification phase.

2.1 Automatic bot identification

Broadly speaking, several approaches have been proposed in the literature for finding different demographics of an author on social media. Content-based, emotion-based, topic-based and deep learning approaches are the main methods proposed by researchers in this field. In this section, we will provide an overview of relevant related works on bot detection. Random forest was the classification technique that has been proven to give the best performance for bot detection on Twitter (Lee et al. 2011; Varol et al. 2017; Singh et al. 2018).

In Fernquist et al. (2018), the authors applied a content-based approach that uses the random forest algorithm to recognize automatic behaviors and detect bots tweeting about the Swedish election. When training their model, they used 140 different features which were divided into two types: metadata and content-based features. The authors found that the most significant feature is produced by the calculation of the number of likes an account has given divided by the number of friends the account has. The second most significant feature they identified is the ratio between the number of followers and friends, followed by the time between retweets. This method achieved good performance. Chu et al. (2012) focus on the classification of human, bot and cyborg accounts on Twitter. They used different features such as those related to account properties, tweeting behavior and tweet content. For the classification phase, they applied the random forest algorithm and they obtained 98% accuracy for humans, 96% for bots and 91% for cyborgs. SentiBot (Dickerson et al. 2014) is a sentiment-aware architecture for identifying bots on Twitter. The authors used a combination of features including tweet syntax, semantics and user behaviors to distinguish human and social bots. To test this method, they employed a large dataset relating to the 2014 Indian election and trained it with several machine learning classifiers. They show that a number of sentiment related factors are keys to the identification of bots, significantly increasing the area under the ROC curve (from 0.65 to 0.73). BotOrNot (Davis et al. 2016) is a publicly available service via the website or via Python or RESTAPIs which aims to compute the botness of a particular user (i.e., the likelihood that an account is a bot). Authors in this paper used more than 1000 features which are grouped into six classes: network, user, friends, temporal, content and sentiment. BotOrNot's classifier used random forest technique and achieved 86% accuracy in detecting bots and humans on Twitter. Oentaryo et al. (2016) profiled Twitter users and classified them into three broad categories: broadcast, consumption and spam bots. They developed a new systematic bot profiling framework using a rich set of numeric, categorical and series features as well as a set of classification models. Empirical studies showed that the diversities of timing patterns for posting activities constitute the key features to effectively identify the behavioral traits of different bot types. Based on the precision results, the authors showed that the random forest classifier outperforms the other machine learning algorithms tested with 84.32%. The authors of Alarifi et al. (2016) collected and manually labeled a dataset of Twitter accounts, including bots, human users and hybrids (i.e., tweets posted by both human and bots). In this work, random forest and Bayesian algorithms reached the best performance at both two-class (bot/human) and three-class classification (bot/human/hybrid). Recently, in Hall et al. (2018) the authors proposed a new strategy for bot detection in Wikidata using a set of comment-based features of user. The comments-based features help to examine the editing behavior of registered and non-registered users. For classification, the authors tested the random forest algorithm and a gradient boosting model and applied optimization by hyperparameter for both classifiers. They showed that this method yields a high level of fitness (ROC-AUC= 0.985). In the work of Cai et al. (2017), a behavior enhanced deep model (BeDM) for bot detection was proposed and applied to detect bots under two deep learning frameworks. This method regards user content as temporal text data instead of plain text and fuses content information and behavior information to capture the latent features. Using a honeypot method, the authors of this study collected a public Twitter-related dataset for their experiments using a convolution neural network (CNN) network. Their model was investigated by varying the number of filters, the filter width and the number of hidden units in the hidden layer. Experiments using tenfold cross-validation showed a result for precision, recall and F1-measure as 88.41%, 86.26% and 87.32%, respectively. In Wei and Nguyen (2019), the authors developed a recurrent neural networks (biLSTM) word embeddings-based model to classify human and spambot accounts on Twitter. This model requires no prior knowledge or handcrafted features. Experiments on the cresci-2017 dataset showed that their proposed approach achieved encouraging performance compared with existing state-of-the-art bot detection systems. They also confirmed the ability of using biLSTM with word embeddings in their model to detect phishing email, webpages or SMSs. Kudugunta and Ferrara (2018) proposed also a deep neural network based on contextual LSTMs architecture that exploits both content and metadata to detect bots on Twitter. There are two classification tasks in this study: account-level bot detection and tweet-level bot detection.

2.2 Automatic gender classification

Several studies have been carried out on the classification of users' gender on social media. In previous studies, many researchers have explored different methods such as content-based, stylometry-based, emotion-based and deep learning models to address this issue. In this section, we will provide an overview of these relevant related works. This can provide a basic understanding on the state-of-the-art of gender classification methods and approaches. In Isbister et al. (2017), the authors proposed a content-based approach to classify the gender of a blog author on five different languages: English, French, Spanish, Swedish and Russian. They used features from Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al. 2001). Experimental results showed that by using support vector machine with features from LIWC, the authors could obtain an accuracy between 73 and 79% depending on the language. In Daneshvar and Inkpen (2018b), authors presented a new content-based approach for gender identification in Twitter. They proposed a linear support vector machine classifier, with different types of word and character n-grams as features. This model was the best-performing model in textual classification, with the accuracy of 82.21%, 82% and 80.9% on the English, Spanish and Arabic datasets, respectively. Fatima et al. (2017) investigated content-based features (word and character n-grams) and 64 stylometry-based features (11 lexical word-based, 47 lexical character-based and 6 vocabulary measures) for the identification of gender and age traits on multilingual corpora. For the classification, several machine learning techniques were used. For gender prediction, best results were obtained using random forest and naive Bayes classifiers. Patra et al. (2018) described a multi-modal author profiling systems using Twitter datasets in three different languages: Arabic, English and Spanish. This system aimed to identify the author's gender. An existing image captioning system was used to extract information present in images. As features, the authors extracted mainly latent semantic analysis, word embeddings and stylistic features from tweets as well as captions. Among three languages, the proposed model for English outperformed the other two languages with an accuracy of 77.37% using support vector machine. Rangel and Rosso (2016), Rangel and Rosso investigated the impact of emotions on gender and age identification. The authors proposed a graph-based approach to model the way people use the language and the emotions when writing. The graph was built with the different parts of speech of user's texts and enriches it with semantic information with the topics they speak about, the type of verbs they use and the emotions they express. This approach that employs both style-based and EmoGraph-based features allowed to identifying gender with an acceptable accuracy using support vector machine classifier. In Safara et al. (2020), Safara et al. used character-based features, syntax-based features, word-based features and structure-based features to detect the gender of an email author. An artificial neural network (ANN) was employed as a classifier, and the whale optimization algorithm (WOA) was used to find optimal weights and biases for improving the accuracy of the ANN classification. Their proposed model achieved an accuracy of 98%. In Garibay et al. (2015), the authors described their methodology for age and gender prediction in Twitter. The authors presented each tweet with a bag of words in a vector space. TF-IDF measure was used to assign a value to each word in a vector. The proposed system presented some failures with the classification of the gender class which affected its performance. To justify this failure, the authors believed that the training of models was over-fitted by the number of estimators in both classification and regression random forest models used. In Takahashi et al. (2018), authors described a neural network model for the gender identification on Twitter named "Text Image Fusion Neural Network (TIFNN)." This method consists of extracting information from written messages and images shared by users. In order to leverage the synergy of the texts and images, the proposed model computes the relationship between them using the direct product. TIFNN achieved accuracies of 84–90% for the different languages used. Basil et al. (2019) tackled the issue of age profiling using convolutional networks. They used datasets collected from Facebook and Twitter. This dataset has been tested the deep convolution neural network algorithm (DCNN) for the classification of age group as a two class such as teenager and adult age group. DCNN had the best performance, reaching a precision of 89% in the validation tests. Another interesting work relying on deep learning models was proposed in Mac Kim et al. (2017) where the authors address gender, age and user type ("individual," "organisation" and "other") profiling on Twitter data. Authors tackled all tasks as a graph vertex classification task using GRNNs based. For this purpose, they employed recursive neural networks and they achieved good performance. In Mabrouk et al. (2020), the authors proposed a new ontology-based profile categorization approach. In order to train a gender predictor systems, experiments using the gender classification dataset 2018 confirm the performance and the effectiveness of this method.

To conclude, several researchers have developed various methods to solve the bots detection problem and the gender identification task. Random forest was the classification technique that has been proven to give the best performance for bot detection on social networks, especially on Twitter. In the next section, we present our motivation for the proposed approach for bot and gender profiling in PAN 2019 (Pardo and Rosso 2019). The main objective of our system is to show how the writing style employed by users in online social media helps us to profile them. More precisely, how can writing style of user differentiate between bots and humans, as well as identify her/his gender.

3 Motivations for proposed approach

The proposed approach solves the problem of the classification of Twitter messages in order to detect bot profiles from human ones. Then, in case of a human, identify her/his gender (man or woman). The strong motivations behind our own work are illustrated in the following points: (1) Bots have become ubiquitous in online social media networks, especially in Twitter during the last 15 years (Dickerson et al. 2014; Davis et al. 2016; Subrahmanian et al. 2016). Therefore, detecting Twitter bots has become an important research domain across the globe for a variety of security purposes. (2) Classification of Twitter messages is especially a challenging problem compared to the classification of long texts such as scientific articles and newspapers (Naouar et al. 2017) that has been the focus of research effort during the last two decades. Indeed, their informal style makes classification hard because there is very little writing style information in short texts (Omri 2004b, a). Therefore, the majority of researchers focus on the content of tweets (Mehrotra et al. 2013; Najib et al. 2015; Nieuwenhuis and Wilkens 2018; López-Monroy et al. 2020) more than using the writing style of users to classify Twitter messages. Therefore, we present our system in order to determine which kinds of stylistics information have the greatest impact to solve that issue. We used a set of new statistical features based on the writing style of the authors. (3) Motivated by the positive results of statistical approaches (Fkih and Omri 2013) and traditional machine learning models for various classification problems (Sendi et al. 2019; Fkih and Omri 2020; Boukhari and Omri 2020), we present in this paper our proposed method-based machine learning. Random forest is one of the most popular supervised machine learning techniques and shows relatively superior generalization performance for many classification tasks in recent years (Surendran et al. 2014; Garibay et al. 2015; Maitra et al. 2016; Kyebambe et al. 2017).

4 Proposed approach for Author Profiling

4.1 General concept of the proposed approach

In this section, we describe our proposed model used to solve the task of bot and gender profiling in PAN 2019 (Pardo and Rosso 2019). This method is designed to identify two types of classes: bot and gender. Bot and gender classification are considered as separated problems. We determinate firstly whether an author is a human or a bot, and then, we predict whether an human is a male or a female.

Figure 1 shows the general process of this study, which is composed of three major parts. The first part is data preprocessing. All operations in this step were done using the Natural Language Toolkit (NLTK). The second part is related to features extraction, which is the major part of our study. We focus on features based on the writing style of authors to solve the present problem. The features are divided into four categories: character-based features, word-based features, syntax-based features and Twitter features. These features have been adopted in previous text classification works and proved effective for online text classification (Flekova et al. 2016; Ashraf et al. 2016; Guimaraes et al. 2017;

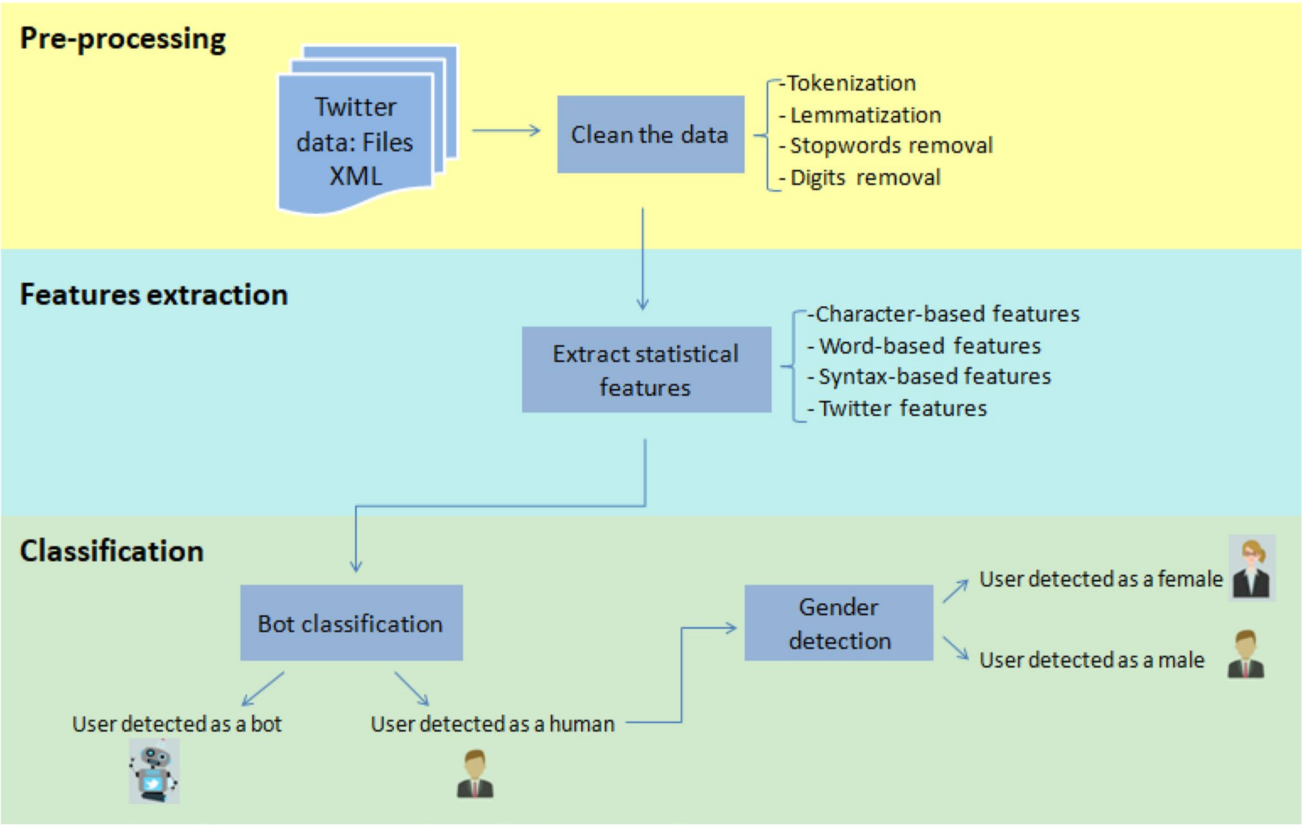


Fig. 1 Overall process of the proposed approach

Safara et al. 2020). The final part of this paper focused on Twitter users classification. This part is composed of two main phases: bots detection and gender classification, i.e., distinguishing bots from humans and in case of a human, identifying his/her gender (male or female). In this part, the prepared features were considered to be an input for different supervised machine learning techniques. We found that the random forest classifier which is frequently used for online text classification outperforms all other algorithms in terms of performance. We note that we used the same concept for bot and gender classification for each respective language. More details about our proposed approach are presented in Sect. 4.2.

4.2 Detailed procedure

4.2.1 Preprocessing and text analysis

To answer the second research question, we describe in this section the different steps we did in the preprocessing phase. The preprocessing of the data is an important step, and its goal is to clean data and to make it ready for applying machine learning classifiers to it. The first step we did in this phase is the concatenation of each user's tweets in order to have only one document for a single author. Then, we perform some transforming tasks of the tweets (tokenization and lemmatization). In addition, we remove digits, we remove also stopwords (such as "the," "and" and "for" in English and "es," "por" and "el" in Spanish) from tweets because they don't provide any useful information for the classification task. There is a default list of stopwords in the NLTK library in several languages, especially English and Spanish. So, in this work, we delete the stopwords provided by NLTK library. It is important to mention that we did not remove punctuation marks, emojis and other special symbols, because these lexical structures are helpful in bot and gender detection and will be used in the present work.

All operations in the preprocessing step were done using NLTK library, a suite of program modules written in Python.

4.2.2 Features extraction

Features extraction and selection from the data are a very critical process for AP task and had a great impact on the experimental results. Previous works on the AP task explore different classes of features for bot and gender detection. To the best of our knowledge and according to the literature, stylometry features are very helpful in distinguishing between tweets written by different categories of people (bots, humans) (Flekova et al. 2016), because online social media users may have different writing styles. They may share texts by using uppercase or lowercase letters, repeating characters or words in one sentence, employing some special symbols and emojis and so on. For example, an important difference between a human profile and a bot profile, that we noticed from the two datasets, is the usage of hashtags (#) and mentions of users (@). We think that bots use to insert hashtags in their texts as a way to increase their reach, and they try to mention multiple users to call their attention. In addition, bots retweet content of others human users to ameliorate their Twitter accounts. Also, they share a high amount of links in their publications compared with human, and this is a typical behavior of spam bots. Considering the above different factors aforementioned, we assume that the statistics of these special tokens in user-generated texts may differentiate bot profiles from human ones, as well as predict the gender of human users. Our proposed approach is purely statistical and based only on tweets from the Twitter users. It accepts input from any document written in English or Spanish, i.e., features used can be applied on any language for bot and gender detection. In our proposed method, we used the same set of features for the bot detection and the gender prediction tasks for the two different languages.

The features used in our study are divided into four categories: character-based features, word-based features, syntax-based features and Twitter features. Each category has its own features as explained in the followings (for details, see Table 1):

Table 1 List of the features used in the development of our proposed system for PAN 2019 bot and gender profiling task

Feature	Description (in terms of: number of)
1 retweet-count	Retweets per tweet (words starting with "rt")
2 mention-count	Mentions per tweet (words starting with @)
3 bot-count	Tweets mentioning the word bot
4 URL-count	URLs per tweet (words starting with http)
5 emojis-count	Emojis per tweet
6 hashtags-count	Hashtags per tweet (words starting with #)
7 characters-count	Characters per tweet
8 lowercases	Lower cases per tweet (a–z)
9 uppercases	Upper cases per tweet (A–Z)
10 spec-char-count	Special characters per tweet
11 words-count	Words per tweet
12 cap-word	Capitalized words per tweet
13 short-words-count	Short words with 3 characters
14 long-words-count	Long words with 6 or more characters
15 hapax-leg-count	Words that only once occur (hapax legomena)
16 hapax-disleg-count	Words occurring twice (hapax dislegomena)
17 punctuation-count	Punctuation marks (? ! ; , ” ” ’ ’ ...)

- Character-based features: number of characters, number of lower cases, number of upper cases.
- Word-based features: number of words, number of short words , number of long words, number of capitalized words, number of words that only once occur, number of words occurring twice.
- Syntax-based features: number of special characters, number of punctuation marks.
- Twitter features: number of retweets, number of mentions, number of tweets mentioning the word bot, number of URLs, number of emojis, number of hashtags.

Tables 2 and 3 show some examples of bots and human tweets extracted from the English and the Spanish datasets, respectively, before the preprocessing step.

For the two languages, we confirmed that bots employed links and hashtags in their posts and shared content of others profiles (retweets) more than humans. To emphasize on the different features we employed in our study, Fig. 2 shows an example of English tweet, written by human, before and after preprocessing.

This is an English tweet written by a male user, containing 40 characters (lower cases), 6 words, 3 long words; punctuation-count is 2, emoji-count is 3; there are no hapax legomena, no hapax dislegomena, no using of the word "bot," no capitalized words, no links, no hashtags, no short words, no upper cases, no retweets, no mentions and no special characters.

4.2.3 Proposed algorithm

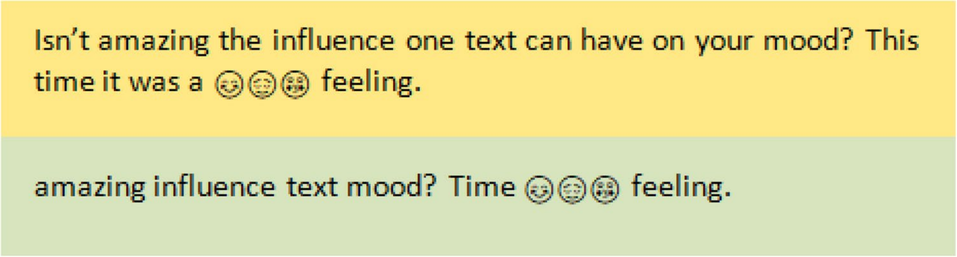
Motivated by the positive results of traditional machine learning models for AP challenge (mentioned in Sect. 2), we explore in this work how these strategies would provide good results in the task at hand, i.e., in bot and gender author profiling task. To this end, our proposed approach relies on applying supervised machine learning algorithm: "Random

Table 2 Some tweets extracted from the English dataset

Bots tweets examples	
1aec68776841385d9e1a6c7d557ca690	Shore Scene with Groups of Figures, 1865 https://t.co/Br8YBxidwK https://t.co/DhqIADup1T
b81abde242dd95f51f3b437e2469fc86	@CityLab https://t.co/jjcLyTTCud
b4a2690000147f44744086b746a0c068	@Sylwiassta thanks for tweeting! Trying to build up awareness
Human tweets examples	
b3ac8a5d2e6663d39d429e34c6a73f12	Fashion Note: Little change in men's clothes this year
1a96a152629a48c236ff10b05188785f	"We're all such products." - Invisible Monsters
1ad7a4b898fff3df9ead777e9ae9e73e	How many generations left until cursive writing is dead?

Table 3 Some tweets extracted from the Spanish dataset

Bots tweets examples	
7b1fcfda1634c2ab7e6f6d3fc4924963	Quince días han tardado... https://t.co/j2bVVpXDJ8
1aaf89b2227df6e550b115797d127a23	RT @eRTme: Atleta sobrevivió a ocho cuchilladas en atraco en cerro de Cali https://t.co/GpEkoy1ws1 https://t.co/1VYLhb8X88
1679d4554c8e9c34a9a037b13b6c5951	Te enseñamos a costumizar tu ropa #en3pasos http://t.co/YhFeRfJJJ4
Human tweets examples	
df3ce515e5b4b36d328c23c86b05b95c	Los nombran en un cargo público y cierran sus cuentas de Twitter o hacen limpieza general de posteos. Nadie resiste un archivo
1824b189d59a8f91f0faa322a575493d	CEESP celebra decisión sobre el Dragon Mart: La decisión del gobierno al no aprobar el proyect.
2221b2c29da141ef1865894a6a5407b9	Cambios en la pizarra del Dólar: Bajó a $32,22 la compra (– $0,06) Bajó a $33,62 la venta (-$0,06)



Isn't amazing the influence one text can have on your mood? This time it was a 😊😌😵 feeling.

amazing influence text mood? Time 😊😌😵 feeling.

Fig. 2 Illustration of text post-processing and preprocessing

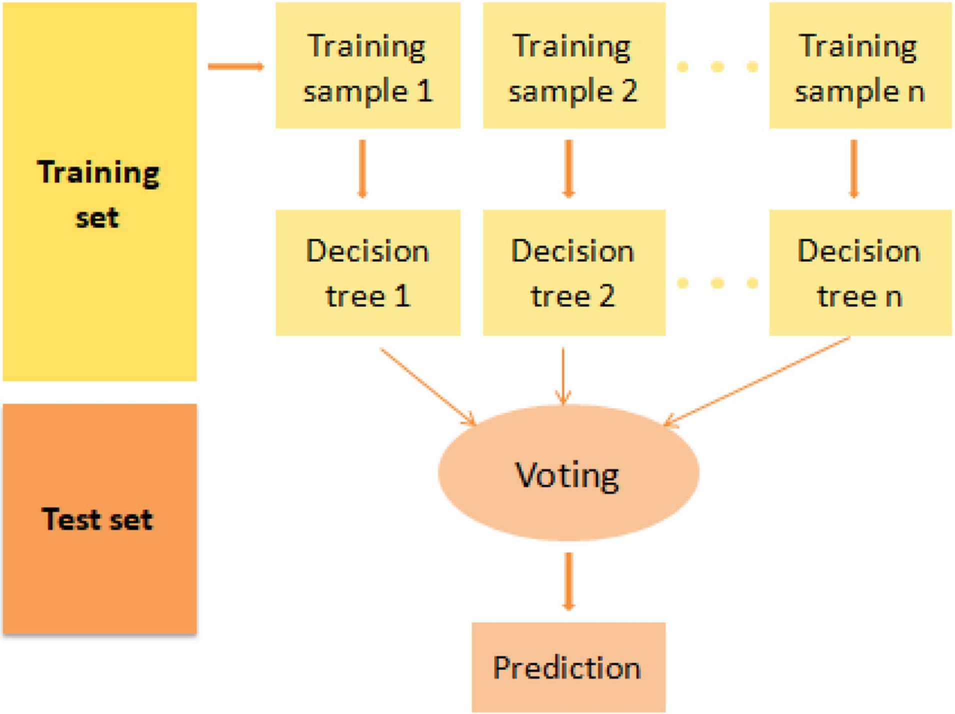


Fig. 3 Workflow of random forest algorithm

Forest.” We chose to train our model with this algorithm because of its best performance in recent years for AP problems and other various domains (Akar and Güngör 2012; Chu et al. 2012; Dong et al. 2013; Inuwa-Dutse et al. 2018; Yang et al. 2019; Li et al. 2020).

Random forest (Breiman 2001) is a supervised learning technique which is widely used for classification problems. In this technique, bootstrapping strategy is employed to partition the set of features into multiple training subsets. Then, the algorithm trains a model (decision tree) on each subset in the training data and gets a prediction result from each decision tree. Then, it performs a vote for each predicted result. Finally, it selects the prediction result with the most votes as the final prediction. Figure 3 shows the general workflow of random forest algorithm. The pseudo-code of our proposed model is shown in Algorithm 1, including the preprocessing phase, the features extraction phase, the step sequence of the random forest algorithm as well as the evaluation metrics which are defined. The random forest algorithm works as follows: For each tree in the forest, we select a bootstrap sample from T where $T^{(i)}$ denotes the ith bootstrap. Then, at each node of the tree, instead of examining all possible feature splits, we randomly select some subset of the features $m \subseteq M$, where M is the set of features. The node then splits on the best feature in m rather than M. In practice, m is so much smaller than M. Here, we define the training set T by:

$$T = \left\{ (x_1, y_1), \dots, (x_n, y_n) \right\} \tag{1}$$

To obtain the best model, it is necessary to adjust the parameters of the algorithm. For random forest, the most important setting, is: n_estimators is the number of trees in the forest. A higher number of trees give better performance but increase the complexity of prediction in terms of time (i.e., it makes the code slower). It should choose as high value as our system can handle because this makes predictions stronger and more stable. Empirically and after some experimental steps, we found that a good setup for both gender and bot classification was: $n\_estimators = 100$.

Algorithm 1: Proposed APSF algorithm

```
Input: Training set T, M features, $n\_estimators$: number of trees in forest
Output: Model with evaluation
begin
    Preprocessing phase
    Extraction of statistical features from the preprocessed data
    M ← Total set of features
    Function RandomForest($T, M$)
        $Z \longleftarrow \emptyset$
        for $i = 1$ to $n\_estimators$ do
            /* select a bootstrap sample from T */
            $T^{(i)} \longleftarrow$ A bootstrap sample from T
            $z_i \longleftarrow RandomizedTreeLearn(T^{(i)}, M)$
            $Z \longleftarrow Z \cup \{z_i\}$
        end
        return Z
        Model ⟵ RandomForest(Z, M)
        Performance ⟵ evaluate(Model)
         Evaluate model with:
        Accuracy calculation /* with equation 2 */
        Precision calculation /* with equation 3 */
        Recall calculation /* with equation 4 */
        F1-score calculation /* with equation 5 */
        out ⟵ (Model, Performance)
        return out
    End Function
    Function RandomizedTreeLearn($T, M$)
        At each node:
        /* $m < M$ */
        m ⟵ the smallest subset of M
        From m, split the best feature
        return The learned tree
    End Function
end
```

It is important to note that, in this study, we rely only on our best performing model (i.e., random forest (RF)). However, we find it important for our next research in this area to train other different models with the same inputs data. So, we examined other machine learning techniques such as support vector machine (SVM), logistic regression (LR) and k-nearest neighbors (KNN). A CNN architecture was also tested in this paper. For both languages and for the two tasks (bot detection and gender identification), the best found configuration of all used classifiers is presented in Sect. 5.2.

5 Experimental study and results analysis

In this section, we present the datasets used to approach the task in question (Sect. 5.1), the strategy of defining the hyperparameters for all machine learning algorithms used (Sect. 5.2), the different parameters used for the performance evaluation (Sect. 5.3). Finally, the experimental results are discussed (Sect 5.4).

5.1 Dataset description

We used the PAN-AP-19 dataset (Pardo and Rosso 2019) to train our proposed system. It consists of tweets from different Twitter users in the form of XML files; each one corresponds to an author and contains 100 unprocessed tweets. This dataset was made available for English and Spanish languages. The English training corpus contains 4.120 author profiles, whereas Spanish training corpus contains 3.000 author profiles. The dataset is perfectly balanced between bots and humans in both languages. The data were previously split in train and dev for both English and Spanish languages, train for the training and dev for the test or

Table 4 Distribution of data in the PAN19-author-profiling-training corpus for bot and gender profiling task

Language	Bots	Male	Female	All authors	All tweets
English	2.060	1.030	1.030	4.120	412.000
Spanish	1.500	750	750	3.000	300.000

Table 5 Confusion matrix for bots detection

	Actual result/ classification	
	Bot	Human
Predictive result/classification		
Bot	TP	FN
Human	FP	TN

Table 6 Confusion matrix for gender classification

	Actual result/ classification	
	Male	Female
Predictive result/classification		
Male	TP	FN
Female	FP	TN

validation phase. Table 4 shows more statistical details about the datasets used.

5.2 Hyperparameters tuning of the algorithms used

The hyperparameters configuration is an important step, because a bad choice of setting can influence our results in terms of performance. At first, the hyperparameter tuning was done by hand. This gives poor findings. So, we opted to explore the use of the grid parameter search. After some experimental steps using the grid parameter search, to find the best hyperparameter configuration for all tested classifiers, we define each parameter as follows:

- For RF classifier: n_estimators=100,
- For LR classifier: C=1e2 and fit_intercept=False,
- For SVM classifier with linear kernel: C=1.0,
- For SVM classifier with RBF kernel: C=1.0 and gamma=0.1,
- For KNN classifier: N_neighbors=500,
- For CNN classifier: Convolutional filter size=64, Convolutional kernel size=4, MaxPooling: pooling size=4, Dropout rate=0.5, Dense layer=256 units, Layers Activation function= ReLu, Optimizer: Adam, Learning Rate: Adaptive (0.001 ⟶ 0.00001), Loss function: Binary crossentropy, Batch Size=32.

5.3 Evaluation metrics

Evaluation is carried out using four standard measures including accuracy, precision, recall and the F1-score. We can note that the greater the precision, the accuracy, the recall and the F1-score are, the better the model performance. The different parameters are defined as follows:

Table 7 Accuracy results on bot and gender detection

Approach	English %		Spanish %	
	Bot	Gender	Bot	Gender
KNN	67.70	65.97	67.21	62.93
LR	80.66	79.72	79.53	73.82
SVM RBF	86.58	85.49	82.58	81.92
SVM Linear	89.41	87.75	89.19	86.55
CNN	89.70	86.50	89.11	84.90
Proposed APSF	92.45	90.36	89.68	88.88

Table 8 Precision results on bot and gender detection

Approach	English %		Spanish %	
	Bot	Gender	Bot	Gender
KNN	69.88	67.3	68.31	63.71
LR	81	80.64	80.93	74
SVM RBF	87.47	86.32	81.59	81.31
SVM Linear	89.25	87.88	88.25	85.94
CNN	89.74	87.05	87.27	84.29
Proposed APSF	92.76	90	89.32	88.67

Accuracy is the ratio of number of correct predictions to the total number of input samples, i.e.,

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FP} + \text{FN}} \tag{2}$$

Precision is the ratio of correct positive instances among the total of the positive instances, i.e.,

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

Recall is the fraction of correct positive instances over the total of all relevant samples and is computed as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{4}$$

F1-score is approximately the harmonic mean between precision and recall measures, i.e.,

$$\text{F1-score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

where TP, TN, FP and FN are true positive rate, true negative rate, false positive rate and false negative rate, respectively. Tables 5 and 6, called confusion matrix, summarize all these measures for bots detection and gender identification, respectively.

Table 9 Recall results on bot and gender detection

Approach	English %		Spanish %	
	Bot	Gender	Bot	Gender
KNN	64.45	62.41	64.47	61.68
LR	80	78.26	77.31	73.67
SVM RBF	85.68	85.04	84.04	82.94
SVM Linear	89.48	88.06	88.97	86.94
CNN	89.79	85.32	88.87	85.05
Proposed APSF	92.09	90.62	89.69	89.26

Table 10 F1-score results on bot and gender detection

Approach	English %		Spanish %	
	Bot	Gender	Bot	Gender
KNN	67.10	64.76	66.33	62.68
LR	80.52	79.43	79.06	73.83
SVM RBF	86.55	86.06	82.80	82.12
SVM Linear	89.36	87.97	89.23	86.43
CNN	89.76	86.17	89.07	84.67
Proposed APSF	92.43	90.31	89.50	88.69

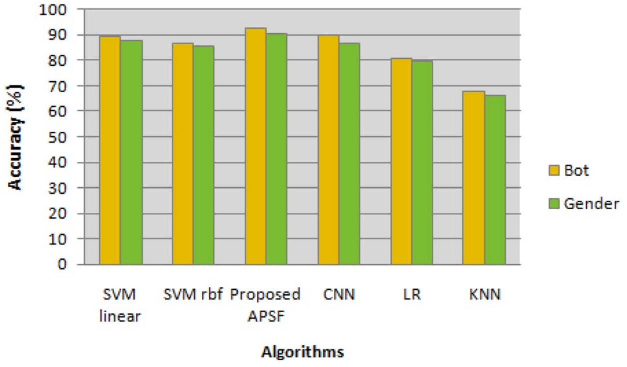


Fig. 4 Comparison of the performances of all models applied in terms of accuracy using English dataset

5.4 Results and discussion

Based on the proposed features, the set of binary classifiers were trained using the English and the Spanish datasets and were applied for bots and gender detection. For each dataset, we performed a tenfold cross-validation on all classifiers. Indeed, cross-validation technique has been shown empirically to yield encouraging results (Daneshvar and Inkpen 2018a). In the subsections below, we first present the performance of our proposal compared to other traditional machine learning algorithms. Second, we compare our approach with other works using the same datasets.

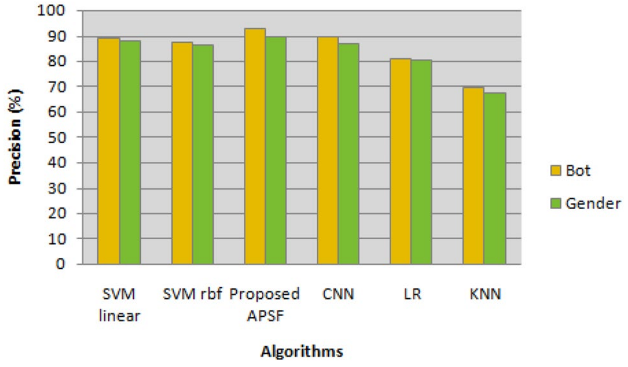


Fig. 5 Comparison of the performances of all models applied in terms of precision using English dataset

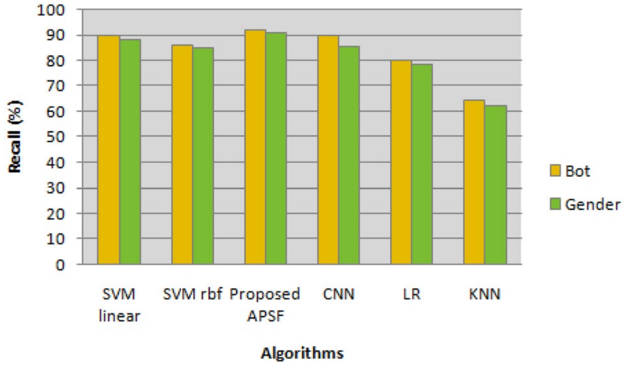


Fig. 6 Comparison of the performances of all models applied in terms of recall using English dataset

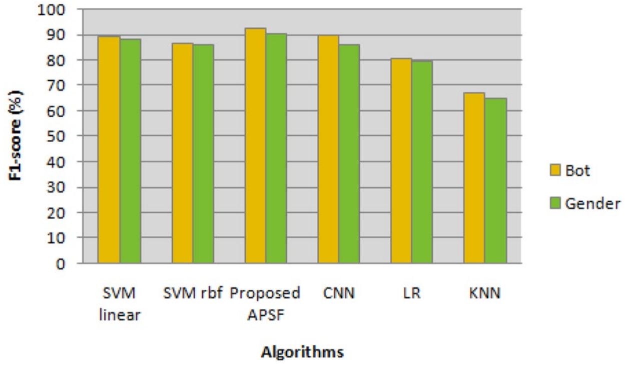


Fig. 7 Comparison of the performances of all models applied in terms of F1-score using English dataset

All of our experiments are conducted on an Acer Aspire *E*5-573 laptop with 2.4 GHz Core i5 5th generation CPU and 16 GB of RAM. For the software environment, we use Windows 10, 64 bit, and a recent version of python environment (Python 3.6).

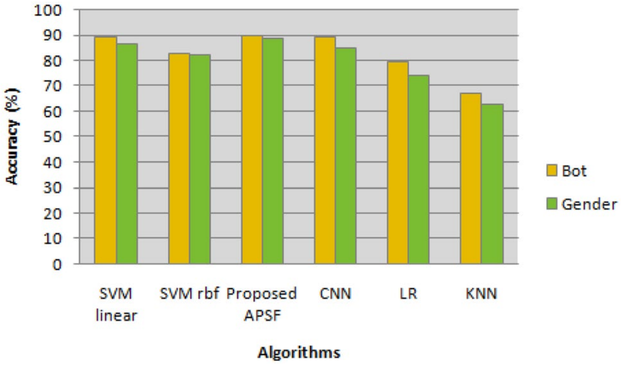


Fig. 8 Comparison of the performances of all models applied in terms of accuracy using Spanish dataset

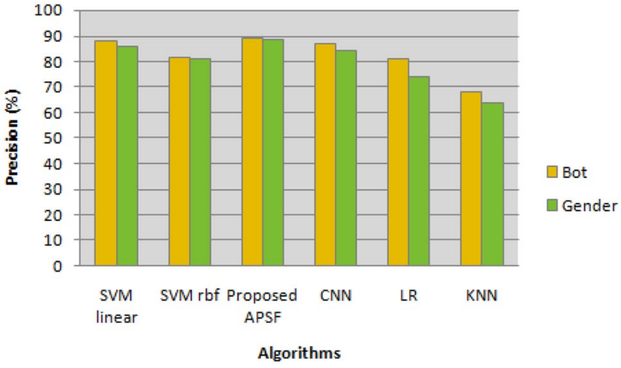


Fig. 9 Comparison of the performances of all models applied in terms of precision using Spanish dataset

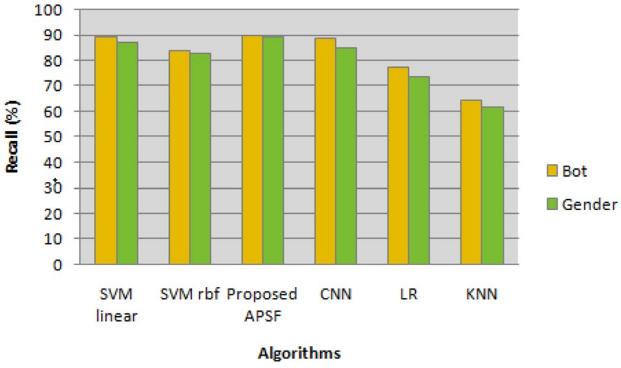


Fig. 10 Comparison of the performances of all models applied in terms of recall using Spanish dataset

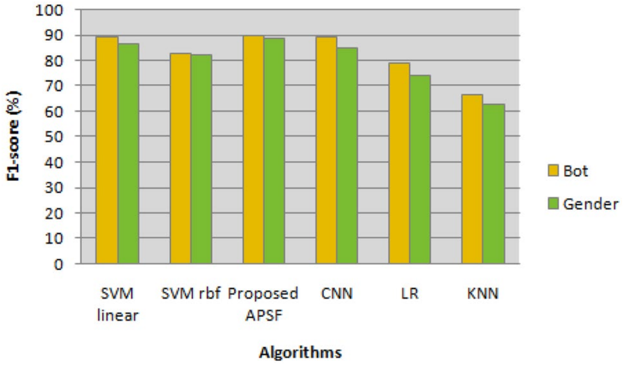


Fig. 11 Comparison of the performances of all models applied in terms of F1-score using Spanish dataset

5.4.1 Comparison against other approaches using the same features

To compare and evaluate our proposed method, data mining techniques such as SVM, LR, KNN and CNN were examined on the same datasets using the same set of features.

After running the random forest algorithm as well as the other supervised machine learning classifiers, Tables 7, 8, 9 and 10 and Figs. 4, 5, 6, 7, 8, 9, 10 and 11 summarize our findings. Indeed, our proposed approach succeeded in finding solution for bots and gender detection in English and Spanish languages using statistical features with the training datasets. As can be seen in results tables, the classification performance of the RF technique (reported in bold) exceeds all other methods results, for both English and Spanish languages, in terms of accuracy, precision, recall and F1-score. Using this supervised classifier, for English language, accuracies of 92.45% and 90.36% are obtained for bot and gender classification tasks, respectively. In terms of recall, bot detection achieved 92.09% and gender classification attained 90.62%. 92.76% and 90% of precision are shown for bot and gender classification tasks, respectively. F1-score of 92.43% was achieved for the bot classification task and F1-score of 90.31% for the gender detection task. On the other hand, for Spanish language, accuracies of 89.68 % and 88.88% are obtained for bot and gender classification tasks, respectively. In terms of recall, bot detection achieved 89.69% and gender classification reached 89.26%. 89.32% and 88.67% of precision are shown for bot and gender detection, respectively. F1-score of 89.5% was achieved for bot detection task and F1-score of 88.69% for gender classification. SVM and LR classifiers also achieved acceptable performance for both English and Spanish datasets (73.82–89.41%). With regard to the CNN algorithm, it also achieved encouraging performance especially for bot identification on the English language (with accuracy of 89.70%). Bad results are obtained using the KNN algorithm.

Graphic comparison shows also that all supervised machine learning algorithms used obtain higher scores when applied for the bot classification than when the same algorithms are applied for gender classification. This difference can be justified by the small training data sizes for gender

Table 11 Comparison of classification results in terms of accuracy of our method with other works provided by Goubin et al. (2019), Puertas et al. (2019) and Ashraf et al. (2019)

Approach	English %		Spanish %	
	Bot	Gender	Bot	Gender
Puertas et al. (2019)	88.07	76.10	80.61	69.44
Goubin et al. (2019)	90.34	83.33	86.78	79.17
Ashraf et al. (2019)	92.27	75.83	88.39	72.61
Proposed APSF	92.45	90.36	89.68	88.88

classification. Also, we can note that our proposed features are more suitable for distinguishing bots from humans than for gender detection. Therefore, the gender classification task is more difficult than the bots detection task. This is likely to happen because bots are likely to insert URLs into their shared tweets and mention others profiles, and humans share texts without more usage of URLs and mention users. Consequently, it makes it easier for the classifiers to distinguish bots from human. Regarding the language of the used datasets, it can be noted that results for English data are higher compared to Spanish, even though same features are extracted for both languages. We can justify this by the complexity of the Spanish data; indeed, Spanish tweets in the datasets of the PAN 2019 bot and gender profiling task (Pardo and Rosso 2019) contain multilingual texts. Therefore, bot and gender detection with Spanish data is more difficult than with English data.

To sum up, according to our results, it makes it easier, for the classifiers applied in our study, to distinguish human from bot in English language.

5.4.2 Comparison against other approaches using the same random forest technique

To assess the efficiency of our approach and demonstrate how does our contribution surpass the existing ones, three state-of-the-art methods proposed to solve the identified task (i.e., the PAN 2019 bot and gender profiling task), that use the same RF technique, were compared with our method.

The baseline systems for the bots and gender profiling task used in our comparative study are presented by: (1) Goubin et al. (2019), (2) Puertas et al. (2019) and (3) Ashraf et al. (2019). This comparison is shown in Table 11 which reports accuracy results. Accuracy was the metric used in these different works, and since we do not have access to the results of precision and recall of other participants, F1-measure could not be calculated. In Goubin et al. (2019), the authors used RF classifier with different features such as number of URL, number of hashtags, number of emojis and number of pronouns in order to predict whether the users were bots or humans. Then, for the gender classification task, they used TF-IDF model based on 1- and 2-word n-grams as features and show that the best performance was reached by the SVM algorithm. In Puertas et al. (2019), for automatic bots identification, the authors used words, number of hashtags, mentions, URLs and emojis as features. For the gender identification subtask, they included the average, kurtosis and asymmetry of the counts of hashtags, mentions, URLs and emojis. They also analyzed the lexical diversity comparing the words used in one tweet to the words used in the rest of the tweets. After some testing on different classifiers, in the case of the English language RF obtained better performance for bots and gender. And for the Spanish language, RF had better accuracy for bots, while LR had better accuracy for genre. Finally, in Ashraf et al. (2019), Ashraf et al. used 27 stylometry features such as number of spaces, number of words and length of tweet for both bots and gender profiling tasks. They tested a range of classifiers including LR, RF, LinearSVC, BernoulliNB, MultinomialNB and SVC. The best results are obtained using RF classifier for both English and Spanish languages. Table 11 summarizes the performance results of all these methods.

As shown in Table 11, the results obtained with our proposed approach are better than the baselines provided by the task organizers mentioned above. Using the same RF algorithm, for the English dataset, our method worked strongly better than others baselines proposed by challenge organizers. Indeed, for bot detection, we achieve an accuracy of 92.45%. Accuracies of 90.34%, 88.07% and 92.27% were achieved, respectively, for the works of Goubin et al. (2019), Puertas et al. (2019) and Ashraf et al. (2019). On the other hand, for the gender classification task, again our approach outperformed all baselines in terms of accuracy (90.36%). Bad result was obtained in the work of Ashraf et al. (2019) with 75.83% of accuracy. For the Spanish dataset, none of the previous works have outperformed the results obtained by our approach using the same RF classifier. For bot classification, challenge organizers reached not very good accuracies (80.61–88.39%). Also, for gender prediction subtask, accuracies obtained by the same organizers were between 69.44% and 79.17%. To sum up, our method outperforms all the baselines and in particular seems more effective especially on English language.

An interesting observation needed to be noted is that all systems previously identified presented by challenge participants (including ours) had higher scores for bot detection than for gender identification. Therefore, we concluded that gender profiling turned out to be much more challenging than bot recognition. In that case, we can explain that this difference in accuracy is justified by the availability of a small English data size compared with the large training corpus used for the bot classification task. It is also worth noting that RF model tends to have a more stable improvements in performance than other baselines when it tested on our well-selected stylistic features. In conclusion, Table 11 confirms the robustness and the efficiency of our method compared to those of the models studied in the literature on the same problem.

6 Limits of the proposed approach

Our random forest-based proposed approach is simple and easy technique for bot and gender identification in Twitter, but not without some limitations. We see that our method presented some failures with the gender classification in Spanish language which affected its performance. The use of the same concepts in both bot and gender classification task can cause problem and affect the performance results. Maybe it was better to use different features for the two different tasks in order to achieve encouraging results to solve this problem. This strategy could be used in future work. Regarding the proposed classifier (i.e., random forest), one of the problems was the computing time. The random forest model requires much more time to train because of its large number of trees generated. The volume of the training data was another important problem in this challenge. Classification problems require a huge dataset to get a good prediction result. Other large Twitter datasets could be examined as well in order to show the effectiveness of the proposed method.

7 Conclusion and future works

In this paper, we present our statistical approach for the PAN 2019 bot and gender profiling task (Pardo and Rosso 2019). Our method consists of two steps. Firstly, we predict whether Twitter accounts are bot or human. Then, we identify the gender of a Twitter human user (i.e., male or female). For this purpose, we used the same features for bot and gender detection (17 stylometry-based features). One of the most significant challenges in this work is the preprocessing of the data. Despite the trivial features used to implement our system, we obtained promising results in both tasks. This demonstrates the importance of the preprocessing phase. For the classification, it was necessary to evaluate different tools with different parameters. We benchmarked our approach with various machine learning algorithms including random forests, support vector machines, logistic regression and k-nearest neighbors. The final experimental results show that our proposed stylometric features work well on such a kind of short and free style messages as Twitter posts. Random forest classifier was the most relevant techniques for both bot and gender detection. In addition, the results of the experiments confirm the robustness and the good performance of our method compared to those of the main models studied in the literature.

There are still issues to explore in this task, and our future work will therefore be structured around four directions. The first direction is to integrate other features in addition of statistical features; for example, we will try to model and extract new deep semantic features from tweets and to implement a hyperparameter optimization to tune the random forest-based proposed model. The second direction consists in using more powerful techniques for features selection such us Boruta algorithm in order to improve model performances. As a third direction, we plan to conduct more tests on other different standard data collections in order to definitively confirm the performance and robustness of our proposed model. The fourth direction is to test our approach with other models based on deep learning techniques in order to give academics and practitioners more amplification on how to deal with the problem of bot and gender profiling.

References

Akar Ö, Güngör O (2012) Classification of multispectral images using random forest algorithm. J Geod Geoinf 1(2):105–112

Alarifi A, Alsaleh M, Al-Salman A (2016) Twitter turing test: identifying social machines. Inf Sci 372:332–346

Ashraf S, Iqbal HR, Nawab RMA (2016) Cross-genre author profile prediction using stylometry-based approach. In: CLEF (Working Notes), Citeseer, pp 992–999

Ashraf S, Javed O, Adeel M, Iqbal H, Nawab RMA (2019) Bots and gender prediction using language independent stylometry-based approach. In: CLEF (Working Notes)

Basil M, Gaikwad S, Salim AS (2019) Deep learning approach based dominant age group based classification for social network. In: International conference on applied computing to support industry: innovation and technology, Springer, pp 148–156

Bessi A, Ferrara E (2016) Social bots distort the 2016 us presidential election online discussion. First Monday 21:7–11

Boukhari K, Omri MN (2020) Approximate matching-based unsupervised document indexing approach: application to biomedical domain. Scientometrics 124:903–924

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Cai C, Li L, Zengi D (2017) Behavior enhanced deep bot detection in social media. In: 2017 IEEE international conference on intelligence and security informatics (ISI). IEEE, pp 128–130

Chu Z, Gianvecchio S, Wang H, Jajodia S (2012) Detecting automation of twitter accounts: Are you a human, bot, or cyborg? IEEE Trans Depend Secure Comput 9(6):811–824

Daneshvar S, Inkpen D (2018a) Gender identification in twitter using n-grams and lsa. In: Proceedings of the ninth international conference of the CLEF Association (CLEF 2018)

Daneshvar S, Inkpen D (2018b) Gender identification in twitter using n-grams and LSA: notebook for PAN at CLEF 2018. In: Working Notes of CLEF 2018-Conference and Labs of the Evaluation Forum, Avignon, France, September 10–14, 2018, CEUR-WS.org, CEUR workshop proceedings, vol 2125

Davis CA, Varol O, Ferrara E, Flammini A, Menczer F (2016) Botornot: a system to evaluate social bots. In: Proceedings of the 25th international conference companion on world wide web, pp 273–274

Dickerson JP, Kagan V, Subrahmanian VS (2014) Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In: 2014 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM 2014). IEEE, pp 620–627

Dong R, Schaal M, O'Mahony MP, Smyth B (2013) Topic extraction from online reviews for classification and recommendation. In: Twenty-third international joint conference on artificial intelligence

Fatima M, Hasan K, Anwar S, Nawab RMA (2017) Multilingual author profiling on Facebook. Inf Process Manag 53(4):886–904

Fernquist J, Kaati L, Schroeder R (2018) Political bots and the Swedish general election. In: 2018 IEEE international conference on intelligence and security informatics (ISI). IEEE, pp 124–129. https://doi.org/10.1007/978-3-319-44564-9_9

Fkih F, Omri MN (2013) Estimation of a priori decision threshold for collocations extraction: an empirical study. Int J Inf Technol Web Eng 8(3):34–49. https://doi.org/10.4018/ijitwe.2013070103

Fkih F, Omri MN (2020) Hidden data states-based complex terminology extraction from textual web data model. Appl Intell 50(6):1813–1831. https://doi.org/10.1007/s10489-019-01568-4

Flekova L, Preoţiuc-Pietro D, Ungar L (2016) 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), pp 313–319

Garibay AP, Camacho-González AT, Fierro-Villaneda RA, Hernandez-Farias I, Buscaldi D, Ruíz IVM (2015) A random forest approach for authorship profiling. In: Working Notes of CLEF 2015—Conference and Labs of the Evaluation Forum, Toulouse, France, September 8–11, 2015, CEUR-WS.org, CEUR Workshop Proceedings, vol 1391. http://ceur-ws.org/Vol-1391/72-CR.pdf

Goubin R, Lefeuvre D, Alhamzeh A, Mitrovic J, Egyed-Zsigmond E, Fossi LG (2019) Bots and gender profiling using a multi-layer architecture. In: Working Notes of CLEF 2019—Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9–12, 2019, CEUR-WS.org, CEUR Workshop Proceedings, vol 2380. http://ceur-ws.org/Vol-2380/paper_235.pdf

Guimaraes RG, Rosa RL, De Gaetano D, Rodriguez DZ, Bressan G (2017) Age groups classification in social network using deep learning. IEEE Access 5:10805–10816

Hall A, Terveen L, Halfaker A (2018) Bot detection in wikidata using behavioral and other informal cues. In: Proceedings of the ACM on human–computer interaction, vol 2 (no CSCW), pp 1–18

Inuwa-Dutse I, Liptrott M, Korkontzelos I (2018) Detection of spam-posting accounts on twitter. Neurocomputing 315:496–511

Isbister T, Kaati L, Cohen K (2017) Gender classification with data independent features in multiple languages. In: 2017 European intelligence and security informatics conference (EISIC). IEEE, pp 54–60

Juola P (2015) Industrial uses for authorship analysis. Math Comput Sci Ind 66:21–25

Kudugunta S, Ferrara E (2018) Deep neural networks for bot detection. Inf Sci 467:312–322

Kyebambe MN, Cheng G, Huang Y, He C, Zhang Z (2017) Forecasting emerging technologies: a supervised learning approach through patent analysis. Technol Forecast Soc Change 125:236–244

Lee K, Eoff B, Caverlee J (2011) Seven months with the devils: a long-term study of content polluters on twitter. In: Proceedings of the international AAAI conference on web and social media, vol 5

Li H, Zhang C, Zhang S, Atkinson PM (2020) Crop classification from full-year fully-polarimetric l-band uavsar time-series using the random forest algorithm. Int J Appl Earth Observ Geoinf 87:66

López-Monroy AP, González FA, Solorio T (2020) Early author profiling on twitter using profile features with multi-resolution. Expert Syst Appl 140:66

Mabrouk O, Hlaoua L, Omri MN (2020) Exploiting ontology information in fuzzy SVM social media profile classification. Appl Intell 66:23

Mac Kim S, Xu Q, Qu L, Wan S, Paris C (2017) Demographic inference on twitter using recursive neural networks. In: Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 2: Short Papers), pp 471–477

Maitra P, Ghosh S, Das D (2016) Authorship verification—an approach based on random forest. arXiv preprint arXiv:160708885

Mehrotra R, Sanner S, Buntine W, Xie L (2013) Improving lda topic models for microblogs via tweet pooling and automatic labeling. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp 889–892

Najib F, Cheema WA, Nawab RMA (2015) Author's traits prediction on twitter data using content based approach. In: Working Notes of CLEF 2015-conference and labs of the evaluation forum, Toulouse, France, September 8–11, 2015, CEUR-WS.org, CEUR workshop proceedings, vol 1391. http://ceur-ws.org/Vol-1391/96-CR.pdf

Naouar F, Hlaoua L, Omri MN (2017) Information retrieval model using uncertain confidence's network. Int J Inf Retrieval Res 7(2):34–50

Nieuwenhuis M, Wilkens J (2018) Twitter text and image gender classification with a logistic regression n-gram model: notebook for PAN at CLEF 2018. In: Working Notes of CLEF 2018-Conference and Labs of the Evaluation Forum, Avignon, France, September 10–14, 2018, CEUR-WS.org, CEUR Workshop Proceedings. vol 2125. http://ceur-ws.org/Vol-2125/paper_183.pdf

Oentaryo RJ, Murdopo A, Prasetyo PK, Lim EP (2016) On profiling bots in social media. In: International conference on social informatics, Springer, pp 92–109

Omri M (2004a) Possibilistic pertinence feedback and semantic networks for goal's extraction. Asian J Inf Technol 3(4):258–265

Omri M (2004b) Relevance feedback for goal's extraction from fuzzy semantic networks. Asian J Inf Technol 3(6):434–440

Pardo FMR, Rosso P (2019) Overview of the 7th author profiling task at PAN 2019: Bots and gender profiling in twitter. In: Working Notes of CLEF 2019-Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9–12, 2019, CEUR-WS.org, CEUR Workshop Proceedings, vol 2380. http://ceur-ws.org/Vol-2380/paper_263.pdf

Pardo FMR, Celli F, Rosso P, Potthast M, Stein B, Daelemans W (2015) Overview of the 3rd author profiling task at PAN 2015. In: Working Notes of CLEF 2015—Conference and Labs of the Evaluation Forum, Toulouse, France, September 8–11, 2015, CEUR-WS.org, CEUR Workshop Proceedings, vol 1391. http://ceur-ws.org/Vol-1391/inv-pap12-CR.pdf

Patra BG, Das KG, Das D (2018) Multimodal author profiling for twitter: Notebook for PAN at CLEF 2018. In: Working Notes of CLEF 2018-conference and labs of the evaluation forum, Avignon, France, September 10–14, 2018, CEUR-WS.org, CEUR Workshop Proceedings, vol 2125. http://ceur-ws.org/Vol-2125/paper_115.pdf

Pennebaker JW, Francis ME, Booth RJ (2001) Linguistic inquiry and word count: Liwc 2001. Lawrence, Mahway, p 71

Puertas E, Moreno-Sandoval LG, Arco F, Alvarado-Valencia JA, Quimbaya AP, López L (2019) Bots and gender profiling on twitter using sociolinguistic features. In: CLEF

Rangel F, Rosso P (2016) On the impact of emotions on author profiling. Inf Process Manag 52(1):73–92. https://doi.org/10.1016/j.ipm.2015.06.003

Rangel F, Rosso P, Koppel M, Stamatatos E, Inches G (2013) Overview of the author profiling task at pan 2013. In: CLEF conference on multilingual and multimodal information access evaluation. CELCT, pp 352–365

Rangel F, Rosso P, Potthast M, Trenkmann M, Stein B, Verhoeven B, Daelemans W et al (2014) Overview of the 2nd author profiling task at pan 2014. In: CEUR workshop proceedings, vol 1180, pp 898–927

Rangel F, Rosso P, Montes-y Gómez M, Potthast M, Stein B (2018) Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. Working Notes Papers of the CLEF

Safara F, Mohammed AS, Potrus MY, Ali S, Tho QT, Souri A, Janenia F, Hosseinzadeh M (2020) An author gender detection method using whale optimization algorithm and artificial neural network. IEEE Access 8:48428–48437. https://doi.org/10.1109/ACCESS.2020.2973509

Sendi M, Omri MN, Abed M (2017) Possibilistic interest discovery from uncertain information in social networks. Intell Data Anal 21(6):1425–1442

Sendi M, Omri MN, Abed M (2019) Discovery and tracking of temporal topics of interest based on belief-function and aging theories. J Ambient Intell Hum Comput 10(9):3409–3425

Singh M, Bansal D, Sofat S (2018) Who is who on twitter-spammer, fake or compromised account? A tool to reveal true identity in real-time. Cybern Syst 49(1):1–25

Subrahmanian V, Azaria A, Durst S, Kagan V, Galstyan A, Lerman K, Zhu L, Ferrara E, Flammini A, Menczer F (2016) The Darpa twitter bot challenge. Computer 49(6):38–46

Surendran K, Gressel G, S T, Hrudya P, Ashok A, Poornachandran P (2014) Ensemble learning approach for author profiling. In: Working Notes for CLEF 2014 conference, Sheffield, UK, September 15–18, 2014, CEUR-WS.org, CEUR Workshop proceedings, vol 1180, pp 1148–1156. http://ceur-ws.org/Vol-1180/CLEF2014wn-Pan-SurendranEt2014.pdf

Takahashi T, Tahara T, Nagatani K, Miura Y, Taniguchi T, Ohkuma T (2018) Text and image synergy with feature cross technique for gender identification: Notebook for PAN at CLEF 2018. In: Working Notes of CLEF 2018-Conference and Labs of the Evaluation Forum, Avignon, France, September 10–14, 2018, CEUR-WS.org, CEUR Workshop proceedings, vol 2125. http://ceur-ws.org/Vol-2125/paper_83.pdf

Varol O, Ferrara E, Davis C, Menczer F, Flammini A (2017) Online human-bot interactions: Detection, estimation, and characterization. In: Proceedings of the international AAAI conference on web and social media, vol 11

Wei F, Nguyen UT (2019) Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings. In: 2019 First IEEE international conference on trust. Privacy and security in intelligent systems and applications (TPS-ISA). IEEE, pp 101–109

Yang KC, Varol O, Davis CA, Ferrara E, Flammini A, Menczer F (2019) Arming the public with artificial intelligence to counter social bots. Hum Behav Emerg Technol 1(1):48–61