Multimedia Tools and Applications (2023) 82:36653–36686
https://doi.org/10.1007/s11042-023-14711-8

A survey of machine learning-based author profiling from texts analysis in social networks

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

Received: 27 August 2021 / Revised: 23 May 2022 / Accepted: 4 February 2023 /
Published online: 18 March 2023

Abstract

Recently, online social networks, such as Twitter, Facebook, LinkedIn, etc., have grown exponentially with a large amount of information. These social networks have huge volumes of data, especially in textual form, which are unstructured and anonymous. This type of data usually leads to cybercrimes like cyberbullying, cyberterrorism, etc. and their analysis has nowadays become a serious challenge. From this perspective and to remedy this topical issue, various techniques have been proposed in the literature. Among the proposed solutions, author profiling represents the newest and most adopted technique by most researchers to discover hidden textual information. The objective of this technique is to identify the demographic or psychological aspects (age, sex, personality, mother tongue, etc.) of an author by examining the text that he has published. In recent years, this area of research has attracted many researchers who seek solutions for potential applications in various fields like marketing, computer forensics, security, etc. Within the scope of this article, we describe the author profiling task. Then, we present a brief thematic taxonomy and an illustration of some profiling solutions from the literature. In particular, different machine and deep learning techniques are detailed and discussed. This work also provides an overview of the main approaches, which we have studied in the literature, highlights the weak points and the strong points of each of these approaches. At the end of this study, a discussion of some research questions is presented and some future directions to circumvent the weaknesses detected in the approaches studied are presented in order to motivate academics and practitioners, who are interested in this problem that we assume essential, to advance solutions for profiling perpetrators on social networks.

✉ 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

Keywords Author profiling · Social networkings · Text analysis · Machine learning · Performance evaluation

1 Introduction

1.1 Context and issues

In recent years, social media networks have grown in popularity thanks to rapid services that help to easily exchange information among people from different geographical areas, ages, gender, and socioeconomic level. The information shared on these online platforms are unstructured and informal. With the increasing amount of this kind of data, that we see every day on the internet, it becomes difficult to know the real identity of the different social network users. For example, in marketing, it is important for a manager to find the demographic aspects (like gender and age group) of the various users who like or dislike their products, with the intention of directing the advertising for exploiting in a better way [12]. In addition, internet has been used to perform fraudulent or illegal acts such as sexual harassment and extortion [20] and other illicit and erroneous acts. Furthermore, fake social media profiles can be seen as a serious threat to user security and the integrity of these platforms. Therefore, designing and implementing effective tools as a solution for this challenging process becomes an unavoidable emergency. The Author Profiling (AP) task is a text classification technique and a subtask of authorship analysis, its goal is to predict demographic and psychological attributes of authors such as age [57], gender [58], personality traits [61], native language [21], political orientation [48], etc. of an author by examining his/her written text. Over the past decade, the AP task has attracted a lot of active research due to its different applications in several fields such as forensics purpose [35], security [29], marketing [61], psychology and terrorism prevention.

Several approaches have been proposed in the literature in an attempt to predict certain personality traits of the authors. Traditionally, there are two types of approaches that have proven to be effective in addressing this identified task: style-based approaches and content-based approaches. The approaches based on style aim to capture an author's writing style using various statistical features including structural, syntactic, and lexical features [6, 23, 70]. On the other hand, content-based methods intend to identify authors' attributes based on the content of their texts. This type of approach is based on topic and semantic structures [3, 51, 73]. We should also mention approaches that combine features of the content-based and style-based methods to enhance their performance [21]. These methods are known as hybrid approaches. The first contribution of the different methods proposed in the literature is based on the extraction and selection of features that can measure the content and the writing style of the author [59]. These approaches aims to construct a features space selected from the text to feed a classification algorithm to determine the author's profile. Many learning algorithms for constructing the classification model have been proposed in the literature, including machine learning algorithms such as support vectors machines [21], random forests [53], decision trees [13], k-nearest neighbors [70], etc., and deep learning algorithms like: convolutional neural network [69], recurrent neural network [32], artificial neural network [65], etc.

1.2 Goals and contributions

The study that we propose in this article positions the problem of AP in social networks, considered large-scale environments in relation to the rapid and diverse evolution of the

quantities of information in their resources. The first and main goal of this research study is to carry out an in-depth review of the AP task, its principle, and its characteristics (with a particular focus on data sources used, features extracted, methodologies and evaluation metrics employed for each method). This article also provides a discussion on some challenges and problems of existing AP approaches and suggests some future research directions for academics and practitioners to advance AP in social media networks.

The main contributions of this work are outlined below:

- ❑ Describe the AP task by presenting its methodology.
- ❑ Propose a new taxonomy of different approaches to author profiling in social networks, and highlights the weaknesses and strengths of each method;
- ❑ Carry out a study of the most recent approaches focused on the problem addressed by drawing up a synthetic assessment according to a certain number of important characteristics to be identified.
- ❑ Provide an overview of the challenges that face the researchers working on this task.
- ❑ Finally, this research work suggests some future directions to address some of those challenges.

1.3 Paper organization

The rest of this article is organized as follows. After introducing the work, in Section 2, we describe the methodology adopted for the collection and selection of the articles studied. Section 3 describes the author profiling task and presents its methodology. Section 4 introduces the proposed taxonomy for author profiling approaches in social media networks. In Section 5, we present and discuss the main techniques used for the AP task. We present and discuss the main techniques used for the AP task. We summarize these methods based on a set of proposed evaluation criteria. Section 6 describe and illustrates the results of the most relevant works. In Section 7, we present a literature synthesis. Section 8 presents the research challenges while Section 9 concludes this work and offers some suggestions for future research.

2 Review methodology

This section presents the methodology we adopted to carry out the following study. In order to succeed in this study, we have structured it around three axes: (i) we provide the different sources of information, (ii) we identify the main search criteria for sources of information allowing us to select the final set. of articles, and (iii) we searched for the relevant questions that we need to answer throughout this study.

2.1 Source of information

We broadly searched for journal and conference research articles as a source of data to extract relevant articles. We used the following databases in our search: Google Scholar,[1]

[1] https://scholar.google.co.in

Fig. 1 Percentage of articles from different types of sources

IEEE Xplore,[2] Springer,[3] ScienceDirect,[4] ACM Digital Library [5]. Also, we screened most of the related high-profile conferences such as ICML, SIGKDD, SKIMA, SIGMOD, ICNC-FSKD, LREC, CLEF, and so on to find out the recent work. In Fig. 1, the percentage of papers reviewed from different types of resources is provided.

2.2 Search criteria

This study was conducted between August 2019 to August 2021. We restricted our research to a period of 12 years. Additionally, we defined two sets of keywords to search the above-mentioned databases since we concentrated on surveying the current state of the art in addition to the challenges and the future direction. In this context, we performed two search iterations. In the first one, we used the following keywords: author profiling in social networks, machine learning for author profiling, text classification, authors classification, features extraction, and features selection. In the second iteration, we tried to look at the related research areas and we used the following keywords: authorship attribution, authorship analysis, authors identification, and user security in social networks.

2.3 Study selection

Based on the used source of information and search criteria, we discovered 1020 articles. On searched articles, we applied a set of selection criteria presented in Table 1 to choose the appropriate research papers. As a first step, we filtered non-ranked articles. After reading the abstract, we excluded some articles that did not meet our criteria. We kept 650 papers. 200 of them are related to the authorship analysis task. However, we chose the most important ones to help us understand our research field. We reviewed the articles completely and only found 50 search papers that represent the studied approaches according to the proposed taxonomy. We used the remaining papers to understand the field, reveal the taxonomy, and propose future directions.

[2] https://ieeexplore.ieee.org
[3] https://link.springer.com
[4] https://www.scopus.com
[5] https://www.acm.org/digital-library

Table 1 Inclusion and exclusion criteria

Inclusion criterion	Exclusion criterion
- Clearly describes author profiling.	- Does not focus on author profiling.
- Peer-reviewed and written in the English language.	- Has common challenges and references.
- Published in reputable journals, conferences, and magazines.	- Articles in a different language than English.
- Written by academic or industrial researchers.	- Short papers, posters, or other kinds of small contribution articles.
- Have a high number of citations in case it is not published in reputable journals, conferences, and magazines.	
- Clearly describes high dimensional data and the different aspects related to it.	
- Latest articles only (last 12 years).	
- In the case of equivalent studies, only the one published in the highest-rated journal or conference is selected to sustain only a high-quality set of articles on which the review is conducted.	
- Articles that propose methodologies, methods, or approaches for author profiling in social networks.	
- Articles that supply methodologies, methods, or approaches for author profiling in social networks.	

2.4 Research questions

The research carried out within the framework of this article aims to answer certain research questions. To reach our objective, we intend to rigorously answer these questions by carrying out a review of existing studies. These questions are summarized in the following points:

– Q1: What are the main reasons and motivations for profiling authors in social networks?
– Q2: What is the methodology used to address the AP task?
– Q3: What methods have been adopted in the profiling of authors?
  This last question can be broken down into five sub-questions which are as follows:
  - q3-1: What types of approaches have been used to solve the AP task?
  - q3-2: What resources and measures were taken into account in the profiling process?
  - q3-3: What classification algorithms were used?
  - q3-4: What are the evaluation metrics used to compare the existing methods?

  And the fourth and final question to explore is:
– Q4: What are the current challenges faced by researchers that should be addressed in the future?

3 Author profiling methodology

AP can be defined as the analysis of human writing in order to find out which classes they belong to, such as gender, age group, occupation, or personality traits. In this section, we

describe the different phases involved in the AP methodology. The AP task consists of four major steps: data collection, pre-processing, feature extraction and selection, and the classification step. The following subsections provide a review of the aforementioned steps.

3.1 Data collection

To address the AP task in social media, the first step to do is "data collection". The data can be collected from many sources such as Twitter, Blogs, Facebook, Instagram, etc. These data collections include texts or documents in English, Arabic, or any other language. In previous works on AP, many researchers [28, 38, 43, 53] have used the PAN dataset (http://pan.webis.de/.); it is a labeled dataset which is provided by the competition organizers. PAN organizers provide participants with training data (texts for which the age, gender, occupation, etc., of the authors, are known) and then evaluate the submitted software on a new unseen dataset. In addition, FIRE (Forum of Information Retrieval Evaluation) has received several methodologies for AP in different languages [56, 63, 68]. Other researchers have manually developed corpora for AP [21, 47, 66, 74].

3.2 Pre-processing

The preprocessing is an essential step for any text classification task, in particular for AP task. Most of the profiles data collected from social networks contain many noisy and missing data, because of the unstructured and informal texts shared on these platforms. Therefore, there is a need to clean the obtained datasets so that the set of features that will be extracted for profiling the authors would produce a good performance result. The goal of the pre-processing phase is to clean data by removing noisy and unwanted data like images, stop words, links, and unnecessary symbols like semicolons, parenthesis, colons, exclamation marks, hashtags, etc. In fact, the presence of this noisy and meaningless data could affect and reduce the results of any analysis [38]. In certain works, this type of features can be useful for the classification step. For example, in [10] the authors use punctuation marks in their study in order to predict the author's profile. Several researchers employed other automatic pre-processing techniques in their studies in order to prepare their data for the analysis phase. Some important pre-processing techniques are: tokenization [28], stemming [15], normalization [40, 53]. Tokenization is the process of dividing the text into small units such as characters, words, phrases, or symbols called tokens. Stemming is the process of transforming terms to their radicals or stems.

3.3 Features extraction

Feature extraction is one of the crucial aspects that is required in solving the AP problem. The features extraction step is aims to extract the needed and significant characteristics from the processed data that will improve the classification performance accuracy. In the AP task, the most used features are based on the style and the content of the text [42]. It is difficult for humans to go through all such text data and find the information of interest and organize a large amount of data. So, various researchers in this domain employed different automatic techniques in order to extract these features. Among these techniques, we mention:

- **Bag of words (BoW):** It is a text representation approach that describes the occurrence of words in a document. As its name suggests, this method does not care about the order of words, it is only concerned with whether known words occur in the text, and

not wherein the document. In [28], Joo and Hwang describe their participation in the PAN 2019 shared task for AP. For each tweet, they extracted n-grams (1 to 3) from the BoW representation. Authors in [42] investigated the role of personal phrases to solve the AP problem and based on all features used in their work, they build a standard BOW representation.

- **Term Frequency – Inverse Document Frequency(TF-IDF):** Many frequently used words can dominate the data, these words can be useless for the model. TF-IDF consists to rescale the frequency of words by how often they appear in all the text. Term Frequency (TF) means the number of times that a word occurs in the document. Inverse Document Frequency (IDF) measures the importance of words in the document. In [9], Basile et al. employed the TF-IDF weighting to extract word n-gram (1 to 2 grams) and character n-grams (3 to 5 grams). For the age classification problem in [11], using the TF-IDF model showed better results than using word2vec representation in the features extraction step. Mabrouk et al. [36] proposed a new approach based on TF-IDF for profiles categorization on Twitter.
- **Word embeddings:** Word embedding is a type of word representation that allows words with similar meanings to have a similar representation. It is capable of capturing the context of a word in a document, semantic and syntactic similarity, relation with other words, etc. Word2Vec representation is the best-known word embedding technique developed by Tomas Mikolov's team at Google [39]. Word2Vec has two neural architectures, called CBOW and Skip-Gram. CBOW receives as input the context of a word (i.e. the terms surrounding it in a sentence) and tries to predict the word corresponding to the context. Skip-Gram does exactly the opposite: it takes a word as input and tries to predict its context. Another popular algorithm is GloVe, developed at Stanford University [49]. Many researchers employed word embeddings techniques to solve the AP task. To address the AP task at PAN 2016, Bayot and Gonçalves [11] used word embeddings and TF-IDF scores. Their results showed that word2vec worked better than TF-IDF for the gender classification task. In [16], the authors presented a combination of stylistic models with word embeddings and used a neural network with GRU activation to predict the gender of the authors. This study was applied to two corpora of two social media varieties: twitter texts and Facebook corpus. For the Twitter dataset, they reached an accuracy of 79%. For the second corpus extracted from Facebook, this method did not show the same performance and obtained an accuracy of 62.1%.

3.4 Features selection and reduction

Sometimes, from one document we extract a lot of features, this can increase the dimensionality of the features space. In this case, many classification algorithms are not able to work with such large features space. So, it is necessary to reduce the features space. Different features selection methods can be used in this step to select the most discriminative features and to remove the redundant or less informative ones. Chi-square metric [24, 52], Information Gain (IG) [43] and Gain Ratio (GR) [21] are commonly used for features selection. The goal of this step is to remove unwanted features from the feature set to give a reduced features vector, and therefore, to predict the author's traits with high accuracy.

3.5 Learning model generation

Once the reduced features vector space is obtained, at this stage, these vectors are inputted to the classifier (a probabilistic model which has the capability to learn and make predictions

on the given data) to obtain the learning model and identify the author of the unknown text. In AP, most researchers used machine learning algorithms and deep learning algorithms as classifiers to generate models for the author's profile prediction. To evaluate the performance of the final model, many techniques are used. Cross-validation is generally the preferred method. The data is randomly divided into "k" equal parts; one of these parts is used for testing and the remaining k-1 part for training. Another technique for evaluating the performance is "Split Validation", where the dataset is usually split into two sets: training data and test data. For example 80% of data for the training phase and 20% for testing, or 50% for training and 50% for the test phase. When the model predicts the output, there are many measures of performance used to evaluate this prediction: accuracy, precision, recall, F1-score, G-mean, etc.

4 Author profiling main approaches

During the last years, various works have been proposed in the literature for author profiling on social media. Based on the type of features extracted from the processed data, we proposed a new taxonomy of the existing AP approaches. We classified these approaches into three main types: style-based approach, content-based approach, and hybrid approach. The approaches based on style aim to predict authors' attributes based on their writing styles. Using this type of approach, researchers employed various statistical features including structural, syntactic, and lexical features. On the other hand, content-based approaches intend to identify authors' aspects based on the content of their texts. This kind of approach is based on topic-based features, semantic structures, BoW representation, etc. The hybrid approach combines features of the content-based and style-based methods. The proposed taxonomy, shown in Fig. 2, will help researchers and academics understand the AP problem and allow them to choose approaches that meet their needs.

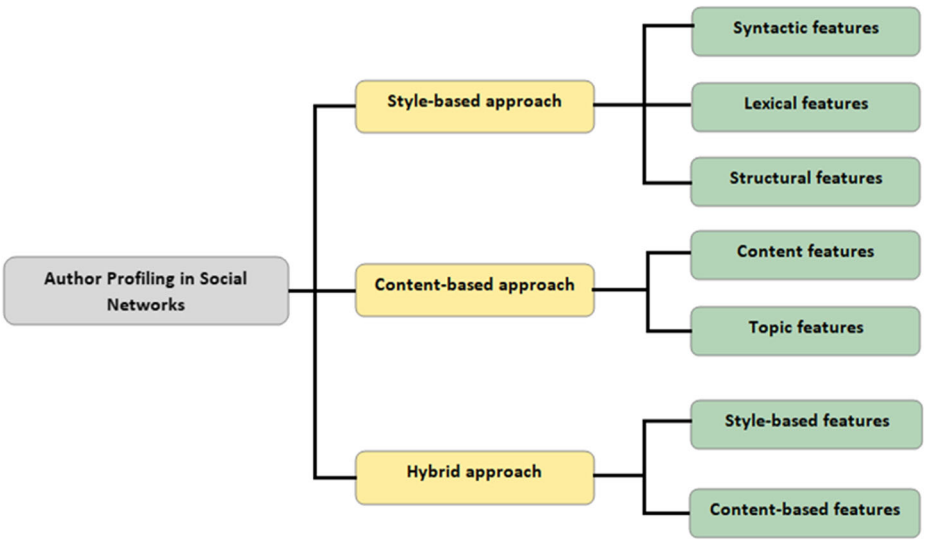

Fig. 2 Proposed author profiling approaches taxonomy

4.1 Style-based approach

Generally speaking, each person has his/her own writing style which can vary depending on gender, age, occupation, geographic localization, etc. The style-based approach uses the personal style to construct the features space to identify the author. These features are namely stylometric features and can be classified into three types: syntactic, lexical and structural features. As an example for syntactic features we can cite: punctuation [10], function words [42], part-of-speech (POS) [1], verbal phrases [16], POS trigrams [14], words per phrase type, etc. Lexical features include content words, frequent words [45], letter frequency, special characters [6], words bigrams, character n-grams [18], word length [23], emoticons [42], etc. The structural features are font color, font size [62], word length distribution and vocabulary richness [6], sentence length [22], URLs, punctuation distribution and word distribution [7], etc.

Various recent works have focused on style-based approaches to predict the demographic characteristics of the authors (such as age, gender, language, personality, etc). For example, in [54] the authors performed experiments for AP on gender and age. They used the PAN-AP-13 corpus in the Spanish language. They considered the stylistic features and the impacts of emotions on gender and age identification. Their approach achieved an accuracy of 63.65% for the gender identification task and 66.24% for the age classification. In 2015, the authors of [50] tackled the AP task at PAN 2015. They used syntactic n-grams as features to predict the author's aspects such as gender, age, and personal traits. This method showed good performance for the Dutch language with an accuracy of 67,98%. Mendoza et al. [42] demonstrated the usefulness of stylistic features in identifying the author of a document. They studied the role of personal phrases for the AP problem on social media. They used words, function words, and POS as features. In their experiments, they examined the PAN-AP-2014 corpus (which contains datasets from blogs, hotel reviews, social media, and Twitter). Their experiments showed that personal phrases reveal more information to identify the gender and age of users on social media. Sandoval et al. [40] examined the PAN 2019 corpus, which consists of English tweets (48335 user profiles with 2181 tweets on average), to predict some demographic traits of celebrities (gender, birth date, degree of fame, occupation) using features based on words, hashtags, mentions, URLs, and emojis. In another study, style-based features were used to predict the age and income of authors. In this context, Flekova et al. [23] built two corpora of tweets (containing 5000 tweets each) to analyze the importance of writing style features in a regression problem. They used a variety of features to capture the language behavior of a user (length of tweets in words and characters, length of words, POS, and number of syllables per sentence, etc.). In their study, they found that stylistic features not only give significant correlations with both age and income but were also predictive of income beyond age. In [53], the authors described their multilingual classification model submitted for the PAN 2019 that is able to recognize bots from humans and women from men on Twitter. They used some style-based features such as words, counts of hashtags, mentions, URLs, and emojis. According to their experiments, they concluded that style-based features demonstrated are very important in distinguishing bots from humans, and the different genders. In [60], Rangel and Rosso proposed a new method to automatically identify the gender and emotions of the authors on Facebook. They chose Facebook comments in the Spanish language as the source of data for their experiments. This method based on stylistic features showed an accuracy of 59% for the gender classification and a recall of 73.7% for the emotions identification task. Recently, in [46], the authors have focused on the AP challenge to know the gender and the age of the authors.

They proposed a new feature selection algorithm based on the weights of some stylistic features. For documents vector representation, a BOW model was employed. Using machine learning as a classification technique, the obtained accuracies were promising. In the same year 2021, Ouni et al. [44] described their method proposed to solve the task of bot and gender profiling at PAN 2019. This method based on the extraction of stylistic features, such as number of URLs, number of words, number of emojis, etc., obtained very encouraging performances.

4.2 Content-based approach

The text consists of words; a word is a sequence of characters; so the order of word or character sequences could provide useful information about the content of the text and the writing style of a particular author. Many researchers have used content-based features to differentiate males, females, different age groups, the country or religion of authors. For example, based on the individual's interests or topics they like to talk about, men mostly used to talk about politics or current events, and sports is the other thing men talk about more. Whereas, women like to talk about shopping, cooking, make up and fashion, also about women's rights. Teenagers like to talk about school and mobile games. Persons in '20-'30s are almost certain to talk about women, love and marriage, or work. Old people prefer to talk about nutrition, pension, and sometimes childhood memories. So, the content of the text is very important to predict the author. Several works have shown the importance of content approaches for the AP task. For example, Cui et al. [17] proposed a new method to classify accounts on Twitter (tweets in April, May, and June 2014: 132.6 million tweets by 23.2 million accounts). They used 11 tweet content features including terms (proportion of tweets with self-reference terms), URLUnique (proportion of unique URLs), etc. Their experiments showed a good classification accuracy. In [43], the authors confirmed that personal phrases presented the essence of texts for the AP task. They considered that the terms located in personal sentences have a particular value and give more information to discriminate the profile of the author. Their approach based on content features showed average improvements of 7.34% and 5.76% for age and gender classification, respectively, when compared to the best results from state-of-the-art (such as the LSA model, LIWC model, SOA model, etc.). Authors in [5] showed the role of the content-based features for the identification of authors personality traits. Anjum and Cheema converted the text into word vectors and counted the frequency of each word. They obtained the best results with this new approach. In [41], Najib et al. described their new proposal to solve the AP task at PAN 2015. They used unigrams with the highest frequencies and the difference in frequencies. The results they achieved are encouraging showing the usefulness of content-based features used. For Spanish gender identification, and accuracy of 84% was obtained. For English age classification, their system achieved an accuracy of 66.9%. And for Dutch personality, they obtained a root mean squared error of 0.124. Kudugunta and Ferrara [34] proposed a new approach to detect whether a given tweet was posted by a human or a bot. They used both content-based features and tweet metadata. The system uses a deep neural network based on a contextual long short-term memory (LSTM) architecture and exhibits the promising performance of over 96% of AUC (area under the curve) to bot detection at the tweet level.

4.3 Hybrid approach

In hybrid approaches, the combination of style and content-based features is used to obtain maximum accuracy of prediction. Many researchers used this type of approach in the

literature. According to the PAN evaluation forums, the most successful work for AP in social media uses a combination of content-based features and style-based features. In 2017, Mehwish et al. [21] have focused on the AP problem on Facebook. They used a set of content-based features (word and character N-grams) and 64 various stylistic-based features (including 11 lexical word based-features, 47 lexical character based-features, and 6 vocabulary richness measures) to predict the age and gender of users. For gender identification, they obtained an accuracy of 87.5%, and for the age identification task an accuracy of 75% was achieved. In [38], Mechti et al. used the English PAN@CLEF 2013 corpus to show the role of stylistic and content-based features in identifying the age of authors. Features used include prepositions, pronouns, determiners, adverbs, verbs, etc. A classification rate of 0.6175 was obtained using advanced bayesian networks. In the research presented by Safara et al. [64] for the author's gender detection of an email author, the features used were divided into four categories: character-based features (like total number of letters, the total number of lower cases, the total number of capital letters, number of characters in a word), syntax-based features (like total number of single quotes, the total number of colons, the total number of periods, total number of commas), word-based features (as total number of words, average length per word, words longer than 6 characters, vocabulary richness), and structure-based features (as total number of phrases, the average number of phrases per paragraph, the total number of lines, the total number of paragraphs). Their model achieved an accuracy of 98%. In previous studies, Joo and Hwang [28] described their participation in the PAN 2019 shared task on AP. They investigated the complementarities of both stylometry and content-based methods to determine whether a tweet's author is a bot or a human, and in the case of humans, identify the author's gender for Spanish and English datasets. Their experimental results demonstrated that the combination of these methods can more precisely capture the author profiles than traditional methods. Kovács et al. [33] also tackled this challenge by extracting semantic and syntactic features from Twitter profiles. They achieved an accuracy of 89.17% for English language tweets in the bot detection task with the AdaBoost technique.

Table 2 shows the difference between the three main approaches based on a set of proposed criteria.

5 Methods of author profiling

For the classification phase and to generate their learned models, researchers used different techniques and methods to solve the AP problem in social media networks. Several probabilistic machine learning and deep learning algorithms were introduced as profiling methods to address the identified task. Some of the most commonly used methods are described and discussed in the subsections below.

5.1 Support vector machine algorithm

Support vector machine (SVM) is a supervised learning technique, it can be used to solve classification and regression problems using data analysis. In AP, SVM is used to predict the different demographic features of authors (classification task). For example, in [73], Yang et al. proposed a Topic Drift Model (TDM) that can monitor the dynamicity of the writing styles and learn the interests of authors simultaneously. They evaluated and compared their approach with the SVM method. According to the experimental results, their model gave the best performance compared with that of SVM. In [21], the authors showed

Table 2 A comparative study of the above three main approaches in AP

Approach	References number	Authors	Importance of text	Number of features	Ambiguity	Addiction
Style-based approach	[54]	[Rangel et al., 2016]	The writing style of the text is important	Around 64 features based on writing style	Becomes ambiguous by comparing the current writing and that after a few years of the same author	On the age of the author
	[50]	[Durán et al., 2015]				
	[42]	[Mendoza et al., 2016]				
	[40]	[Sandoval et al., 2019]				
	[23]	[Flekova et al., 2016]				
	[53]	[Puertas et al., 2015]				
	[52]	[Prasad et al., 2015]				
	[73]	[Yang et al., 2018]				
Content-based approach	[17]	[Cui et al., 2017]	The content of the text is important	Each word is a feature	Becomes ambiguous when several authors write on the same topic	On the psychology and on the mental state of the author
	[2]	[Akimushkin et al., 2018]				
	[43]	[Mendoza et al., 2018]				
	[5]	[Anjum et al., 2018]				
	[41]	[Najib et al., 2015]				
	[67]	[Sendi et al., 2019]				
Hybrid approach	[21]	[Fatima et al., 2017]	Both content and writing style of the text are important	64 stylistic features and each word is a feature	Becomes ambiguous by comparing topics and writing style of the author	On the age, psychology and on the mental state of the author
	[38]	[Mechti et al., 2016]				
	[64]	[Safara et al., 2020]				
	[28]	[Joo et al., 2018]				
	[33]	[Kovács et al., 2019]				

their system working on the AP task for multilingual text composed of English and Roman Urdu, in order to identify gender and age. They focused on AP on Facebook. Their extensive empirical evaluation showed that content-based methods (using word and character n-grams features) outperformed stylistic-based methods (using 11 lexical word-based features, 47 lexical character-based features, and 6 vocabulary richness measures) for both gender and age identification tasks by using the SVM algorithm. In [42], the authors examined the role of personal phrases for the AP task to predict the age and gender of authors on social media. To classify documents, they used the SVM algorithm and they obtained encouraging performances. In [4], the authors focused on both age and gender identification on Twitter by using the visual modality. The authors of this paper aimed to evaluate the pertinence of using visual information to solve the AP task. To classify the tweets, they used the SVM technique using LibLinear. In [50], the authors addressed the AP task at PAN 2015. The method used a supervised machine learning approach (SVM), where a classifier is trained independently for each label (gender and age). Mabrouk et al. [37] proposed a new approach based on TF-IDF for microblog profile categorization. They employed SVM as a machine learning method, and they obtained encouraging results in terms of performance.

5.2 Random forest algorithm

Random forest (RF) is a supervised learning model which is used for classification problems. A forest is made up of trees and more trees mean a more robust forest. RF uses the prediction of each tree to get a more accurate prediction. Various studies related to the AP problem employed RF as a classifier for documents. For example, in [53] the authors presented an analysis of different sociolinguistic features to show how different linguistic characteristics can determine whether the author of a Twitter account is a bot or a human and, in the case of humans, identify the gender of the author. For the classification, the authors analyzed different algorithms. They showed that for the English dataset, RF offered the best performance for bots and gender prediction tasks (macro-F1 score of 91% and 84% for the bot and the gender classification, respectively). Using the Spanish tweets, RF has also achieved better accuracy for the bot classification task (macro-F1 score= 84%). Ashraf et al. [6] presented a stylometry-based approach to identify two author traits (gender and age). The proposed system was trained using different machine learning algorithms including RF. Promising results were obtained on the training dataset (an accuracy of 98.3% for age, 78.7% for gender). In [27], the authors presented their submission to the PAN 2019 bots and gender profiling task. In this work, they proposed a supervised approach using the RF algorithm. They obtained highly competitive bot and gender classification accuracies on English data (96% and 84%, respectively). For the Spanish dataset, they also achieved acceptable performance for the bot and gender identification (88% and 73%, respectively). However, in [45] the authors addressed the AP problem at PAN 2015. The methodology used the RF technique for classification and regression. Their approach presented some failures with the classification of the gender class which affected performance.

5.3 Naive Bayes algorithm

The naive Bayesian (NB) classification method is a supervised machine learning technique that aims to classify a set of data according to some of its properties. This algorithm must first be trained on a training dataset that shows the desired output according to the inputs. During the training phase, the algorithm develops its classification rules using this dataset.

Therefore, these rules will be applied for classifying the data set (test phase). The NB classifier indicates that the classes of the training dataset are known, hence the supervised nature of the tool. There are several categories of this type of machine learning such as Gaussian Naïve Bayes, Multinomial Naïve Bayes and Bernoulli Naïve Bayes. For the AP problem, various studies based on NB were presented in the literature. In [38], the authors proposed a new method for the AP of anonymous English texts (blog posts). This method used the NB algorithm for age prediction. A good classification rate of 0.6175 was obtained. Recently, Gamallo and Almatarneh [25] presented a classification method for bot detection on Twitter. They used the NB technique with features including specific content of tweets and automatically built lexicons. They reached an accuracy of 81% using the English test dataset. However, in the work described in [72], using NB as a classifier, the authors did not achieve high accuracy: 39% accuracy for blogs, 31% for hotel reviews, and 35% for social media. This result was poor and showed that this classifier is ineffective to solve the identified task.

The strengths and weaknesses of each machine learning technique mentioned above are presented in Table 3.

5.4 Deep learning-based author profiling

In the last few years, deep learning methods have also dominated the state of the art and gained tremendous popularity because of their results across the board, natural language processing inclusive, was employed for the first time in 2016 for AP problems. Deep learning (DL) is a subset of machine learning methods, it uses deep neural networks to identify structures in huge volumes of data.

DL models make use of several algorithms such as Multilayer Perceptron Neural Network (MLPNN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), etc. Recently, several works have made efforts to solve the AP task with DL approaches. In [69], the authors described their submission to the PAN 2017 AP shared task (the corpus contains tweets in four different languages: English, Spanish, Portuguese and Arabic). They trained two models for gender and language variety using a CNN architecture, and achieved encouraging performance results. In the same year 2017, Kodiyan et al. [32] presented a new method to predict the gender and language variety of Twitter profiles. Their approach consists of a bidirectional RNN implemented with a Gated Recurrent Unit (GRU) combined with an attention mechanism [8]. Word embeddings were used as features. They obtained an average accuracy over all languages of 75,31% for the gender identification task and 85,22% for the language variety classification task. In 2018, in [31] the authors focused on the Lithuanian AP task for both age and gender identification. They used two DL methods: LSTM and CNN. Comparing their models with the traditional machine learning methods, the DL model is not the best solution for the AP task. In [71], the authors proposed a new approach called "Text Image Fusion Neural Network (TIFNN)" for the gender identification task on Twitter. This solution aims to extract information from written messages and images shared by users. The authors applied DL method to join text and image information. They used CNN for texts and ImageNet-based CNN for images, and they achieved an accuracy equal to 85%. In [26], the authors proposed the CheckerOrSpreader model which aims to differentiate between users that tend to share fake news (spreaders) and those that tend to check the factuality of articles (checkers). This new model is based on the CNN technique and combines word embeddings with features that represent users' personality traits and linguistic patterns used in their tweets. Experimental results showed that the CheckerOrSpreader model achieved acceptable performance (59% of accuracy). In [19], the authors described their approach for bot and gender detection on

Table 3 Summary table of machine learning models used in AP

Classifiers	References Number	Authors	Advantages	Disadvantages
Support Vector Machine	[73]	[Yang et al., 2018]	• Work well with high dimensional data	• Takes long training time for large data
	[21]	[Fatima et al., 2017]	• Uses very less memory	• Do not work well with overlapping classes
	[42]	[Mendoza et al., 2016]	• Offers a great accuracy	• Choosing an efficient kernel function is difficult
	[4]	[Carmona et al., 2018]	• Robust against overfitting problems	• Memory complexity
	[50]	[Durán et al., 2015]		
	[37]	[Mabrouk et al., 2018]		
	[52]	[Prasad et al., 2015]		
Random Forest	[45]	[Garibay et al., 2015]	• Flexible and possess a high accuracy	• Takes long training time compared with other algorithms
	[53]	[Puertas et al., 2015]	• Efficiently run on large data classes	• Overfits to some data
	[6]	[Ashraf et al., 2016]	• Not require preparation and preprocessing of the input data	• Complex
	[27]	[Johansson, 2019]		• More trees in forest increases time complexity in the prediction step
	[10]	[Basti et al., 2019]		
Naive Bayes	[38]	[Mechti et al., 2016]	• Requires less training data	• known as a bad estimator
	[25]	[Gamallo et al., 2019]	• It works very well with text data	• A strong assumption about the shape of the data distribution
	[72]	[Román et al., 2014]	• Fast and easy to implement	
	[21]	[Fatima et al., 2017]		
	[13]	[Bilal et al., 2016]		

Twitter. They employed CNN and RNN techniques based on character and word n-gram models alike. The proposed method "CNN+RNN" reached acceptable performance for the bot detection task (82%-84%), while for gender profiling, the scores obtained were lower (58%-65%).

The benefits and the drawbacks of the different DL methods are presented in Table 4.

In Table 5, we try to summarize the existing deep learning-based approaches and to study its performance according to a proposed set of evaluation criteria. First, we identify which type of features was used for each work. Then, some criteria that indicate the different aspects related to the performance evaluation are provided. These criteria include effectiveness, in addition to the different issues that impact the performance, such as big data handling, overfitting, and hyperparameters tuning.

- **Style-based:** indicates that the proposed approach used style-based features.
- **Content-based:** indicates that the proposed approach used content-based features.
- **Effectiveness:** indicates the capability of the model to achieve the intended findings.
- **Handling big data size:** indicates if the model can deal with very large datasets.
- **Hyperparameters tuning:** indicates if the model requires more hyperparameter tuning. A model that requires a lot of hyperparameter tuning is difficult to implement.
- **Overfitting:** indicates if the model can deal with danger of overfitting.

6 Analysis and discussion

After synthesizing some reference papers, this section is devoted to illustrating and discussing the results of the most relevant works to show how AP performs at different levels according to the proposed taxonomies presented in previous sections. Indeed, several factors can affect the performance findings of the existing approaches. For example, the presence of noisy and unwanted data could affect and reduce the results of any analysis. Other crucial factors regarding features extraction, some researchers have manually extracted features from data to predict the author's traits [10]. The availability of small training data sizes can also affect the precision of proposed models.

Table 4 Advantages and disadvantages of deep learning methods

Advantages	Disadvantages
• Robustness to natural variations in the data is automatically learned	• It is extremely expensive to train due to complex data models. Moreover DL requires expensive GPUs and hundreds of machines. This increases cost to the users.
• The same neural network based approach can be applied to many different data types	• Need abundant data
• It delivers better performance results when amount of data are huge	• It is hard to describe, and is not completely understood.
• Can deal with complex input-output mappings	• Is extremely computationally expensive to train.
• Can easily handle online learning (It makes it very easy to re-train the model when newer data becomes available.)	• Finding an efficient architecture and structure is still the main challenge of this technique
• Parallel processing capability (It can perform more than one job at the same time)	• Patchy support for pretrained models

Table 5 Summary of the main methods based on deep learning

Approach	Style-based	Content-based	Effectiveness	Big data	Hyperparameter	Overfitting prevention
Sierra et al. [69]	x	✓	✓	x	✓	x
Kodiyan et al. [32]	x	✓	x	x	✓	✓
Dzikicne et al. [31]	x	✓	x	✓	✓	x
Takahashi et al. [71]	x	✓	✓	x	✓	x
Giachanou et al. [26]	✓	✓	✓	x	✓	x
Dias et al. [19]	x	✓	x	✓	x	x

From the existing works related to the AP field, we present in this part the most important study in terms of performance. Therefore, this section discusses the work of [64]. The main idea was to predict the gender of an email author using an artificial neural network (ANN) as a classifier and the whale optimization algorithm (WOA) to find optimal weights and biases for improving the accuracy of the ANN classification. This proposed approach was a hybrid that used content and style-based features. In the following subsections, we present in detail the characteristics of the dataset, the features used, and the results of the experimental evaluation.

6.1 Data source and features

In this subsection, we present the characteristics of the dataset used and the features extracted from the data in the work of Safara et al. [64]. To detect the gender of an email author, the authors presented a new approach "ANN-WOA". For their experiment, they used the Enron dataset. Enron dataset is an email data collection and was originally made public and published on the web, by the Federal Energy Regulatory Commission. This dataset contains around 500000 messages. From each message, the authors extracted 48 linguistic features. These features are divided into four categories: character-based features (such as the total number of letters, the total number of lower cases, the total number of characters in a word, the total number of upper cases), word-based features (such as the total number of words, average length per word, words longer than 6 characters, vocabulary richness), structure-based features (such as total number of sentences, total number of lines, total number of paragraphs, average number of sentences per paragraph), and syntax-based features (such as total number of single quotes, the total number of periods, the total number of commas, the total number of colons). To implement this method, the authors used 70% from the dataset for the training phase and 30% for the test phase (more details are shown in Table 6).

Table 6 Characteristics of the Enron dataset used in [64]

Dataset	Enron
Language	English
# of documents	500.000
# of users	160
% Used for training/test	70% for training and 30% for test

6.2 Evaluation metrics

The evaluation phase is very crucial in any classification problem in order to test the performance of the proposed model. To evaluate and compare their proposed method, Safara et al. [64] examined traditional machine learning techniques such as SVM, NB, ANN, and DT on the same dataset. For this purpose, three standard measures including precision, accuracy, and recall using a 20-fold cross-validation technique are used. For more evaluation, we discuss the F1-score and G-mean measurements.

Let TP, TN, FP, and FN be true-positive rate, true-negative rate, false-positive rate, and false-negative rate, respectively. Table 7 called, confusion matrix, summarizes all these parameters.

The different parameters are defined as follows:

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

$$Accuracy = \frac{TN + TP}{TN + TP + FP + FN} \tag{1}$$

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

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

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

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

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

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

G-mean i.e. Geometric Mean measures the balance between classification performances on both the majority and minority classes, and is computed as:

$$G - mean = \sqrt{\frac{TN}{TN + FP} * \frac{TP}{TP + FN}} \tag{5}$$

6.3 Results of the experimental evaluation

The performance results of the new approach proposed by Safara et al. [64] are presented in Table 8. The precision, accuracy, recall, and the F1-measures of all methods using 20-fold cross-validation are shown respectively in Figs. 3, 4, 5 and 6. In Fig. 7 we present the G-mean, which is a measure that tries to maximize the accuracy of the model training. To clarify, a low G-mean is an indication of poor performance in the classification of the positive cases even if the negative cases are correctly classified.

Table 7 Confusion matrix

	Actual class	
	Male	Female
Predicted class		
Male	True Positives (TP)	False Positives (FP)
Female	False Negatives (FN)	True Negatives (TN)

Table 8 Classification results of ANN-WOA approach [64]

Performance classes	Value (%)
Precision	97.16
Accuracy	98
Recall	99.67
F1-score	98.13
G-mean	98.5

As shown in Table 8, the ANN-WOA model [64] achieved high classification performances. First, as we mentioned in previous sections feature extraction is a very important step in the AP task. So, according to the findings of the ANN-WOA model, using both style and content-based features can help to obtain good performances. In addition, the experimental results showed that the WOA algorithm is well merged with the ANN classifier to achieve the best accuracy. Also as illustrated in Figs. 3, 4, 5, 6 and 7, ANN-WOA method outperformed the other machine learning techniques (NB, DT, ANN and SVM) in terms of all performances classes. In terms of precision, this approach was higher than the other machine learning methods examined and it achieved a good value of 97.16% as illustrated in Fig. 3. In terms of accuracy, Fig. 4 shows that the ANN-WOA method outperformed the other classifiers and gives a significant value of 98%. The recall measure is illustrated in Fig. 5, the proposed method also achieved an important value of 99.67%. In terms of F1-score and G-mean, Figs. 6 and 7 show that the model proposed by Safara et al. [64] gives the best measures and were 98.13% and 98.5%, respectively.

7 Literature synthesis

This section is devoted to synthesizing the literature work discussed in this work. It is presented as a table to better summarize this article. Table 9 summarizes the different papers surveyed in the AP field using text analysis approaches. In this table, we present the datasets used, the type of each approach, the set of features selected, the different machine learning techniques employed as classifiers for each experimental study, the obtained results, and

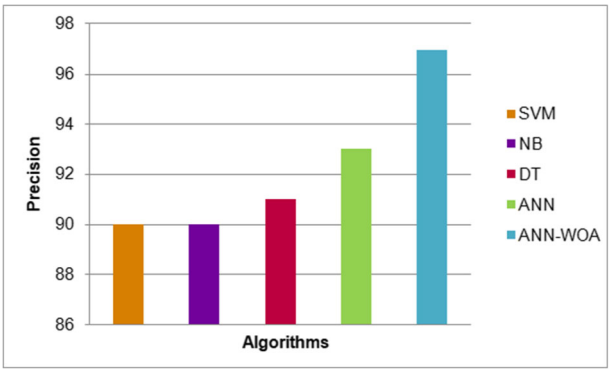

Fig. 3 Comparison of the performances of all methods in terms of Precision [64]

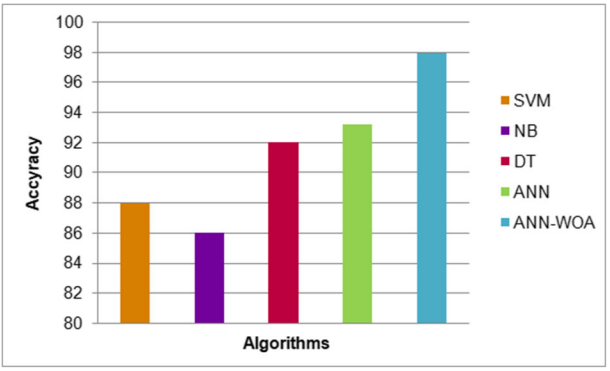

Fig. 4 Comparison of the performances of all methods in terms of Accuracy [64]

the main conclusions of each work. In addition, to better present the literature review, we discuss in this table the performance metrics used to evaluate each approach.

8 Research challenges

Through our extensive work in this survey, we carefully examined several papers based on the AP in social networks and presented a deep-diving analysis of these articles. This work has been summarized in different tables after discussing the main aspects relating to this domain as illustrated in the proposed taxonomies. The aim of this section is to highlight the challenges encountered by researchers in the AP field. We discuss in the following subsections the major challenges inherent in the profiling of the authors on social media networks.

Fake profile: Social media users often spread false information with or without bad intentions. However, using this information in determining the author profile will lead to a fake profile creation (for example profile with false gender or false age). Therefore, there is a

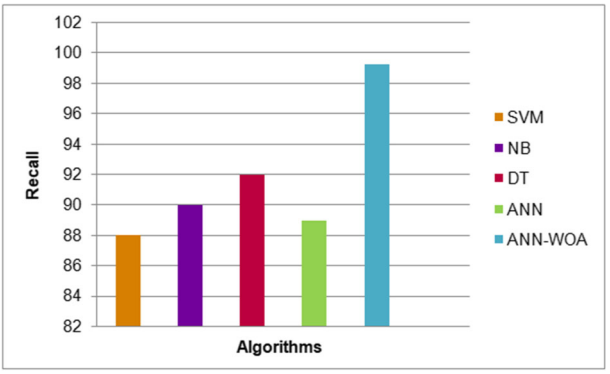

Fig. 5 Comparison of the performances of all methods in terms of Recall [64]

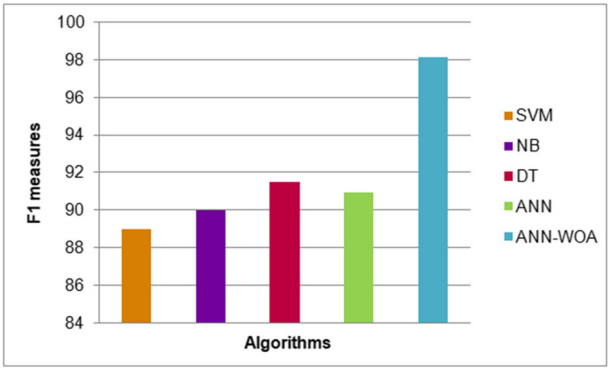

Fig. 6 Comparison of the performances of all methods in terms of F1-score [64]

need for more investigation on the AP approaches that will be able to distinguish fake profiles from true or authentic profiles. This interesting challenge was examined in the AP task at PAN-2019 [55], but all methods proposed to address this problem were not able to provide an efficient fake profile detection model. Indeed, fake social media profiles can be seen as a serious threat to user security and the integrity of these platforms.

Manual techniques problem: In different studies related to the AP task, some researchers tried to extract features manually from the textual data [10], [30]. For example, in [10], Basti et al. proposed a new approach to determining the age and the gender of users on Twitter. In their study, they manually grouped terms belonging to the same class of proposed attributes. They also manually extracted semantic features. This is difficult to implement, requires a lot of time, and cannot be extensively used for classification problems. Automatic techniques for building a features vector space are more efficient and reliable. This problem is very common in content-based approaches using topic-based features.

Monolingual text: The majority of existing AP corpora are developed and available in English and other European languages such as Spanish, Dutch, Italian, etc. and these are monolingual. Monolingual corpus means texts written in just one language. So, researchers

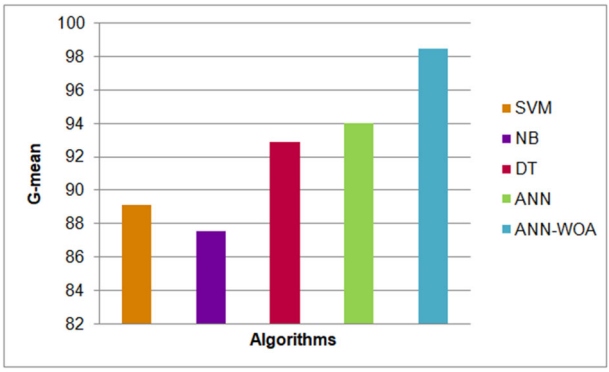

Fig. 7 Comparison of the performances of all methods in terms of G-mean [64]

Table 9 Summary of the proposed author profiling models in social media

Dataset	Approach	Reference	Features	Methods	Obtained results	Conclusion
Twitter	Style-based	Rangel et al. [54]	• Emoticons	SVM	• Acc_g= 63.65%	• Graph-based features are useful for identifying the gender of social media users.
			• Stylistic		• Acc_a= 66.24%	• Gender identification is a more difficult task than age identification.
	Style-based	Duran et al. [50]	• N-grams	SVM	• English:	• Syntactic n-gram with specific tweet features are useful for personal traits prediction than for age and gender classification.
			• Hashtags		Acc_g= 59.15%	• The performance of the model was affected by the use of external syntactic parser.
			• Retweets		Acc_a= 58.45 %	
			• Emoticons		RMSE= 0.1882	
			• URLs			
	Style-based	Sandoval et al. [40]	• Hashtags	• NB	• Acc_o= 57%	• Sociolinguistic features are helpful in celebrity profiling on Twitter.
			• Mentions	• RF	• Acc_f= 65%	• The imbalanced training dataset used was a big challenge, that should be soled by performing oversampling.
			• Stylistic	• LR	• Acc_g= 88%	
			• Words		• Acc_{by}= 37%	
			• Emoji			
			• URLs			
	Style-based	Flekova et al. [23]	• Syllables	SVM	PC=0.352	• Stylistic features give significant correlations.

Table 9 (continued)

Dataset	Approach	Reference	Features	Methods	Obtained results	Conclusion
			• characters			• Temporal stylistic patterns can improve the results results in user socio-demographic predictions.
			• Words			
			• POS			
	Style-based	Puertas et al. [53]	• Hashtags	• RF	• English:	• The features used are very useful for distinguishing bots from humans, and differentiating males from females.
			• Mentions	• NB	$F1_b$=91%	• RF and LR were the most relevant techniques for this challenge.
			• Lexical	• LR	$F1_g$= 84%	
			• Words		• Spanish:	
			• URLs		$F1_b$= 84%	
			• Emoji		$F1_g$= 80%	
	Content-based	Cui et al. [17]	• Term	• GBMDS	• Acc=90.75%	• CDS is a novel learning scheme for Twitter account classification that does not require intensive manual labelling.
			• URL	• DeepDS		• CDS improves the classification accuracy of distant supervision when the heuristic labels are of low quality.
				• CDS		
				• DS+		
				• DS		

Table 9 (continued)

Dataset	Approach	Reference	Features	Methods	Obtained results	Conclusion
	Content-based	Mendoza et al. [43]	• Sentences	• DPP-EXPEI	• Acc_g= 78.88%	• For age profiling, the term selection method was very important.
			• Term	• MNB	• Acc_a= 61.44%	• For gender profiling, the term weighting method was more relevant.
				• KNN		
				• ANN		
	Content-based	Najib et al. [41]	• Content	SVM	• Acc=61.3%	• Content based features are useful in predicting the author's traits.
			• Words		• RMSE=0.196	• This model worked well on the training data than on the testing data.
	Hybrid-based	Joo et al. [28]	• Punctuation	BERT	• Acc_b= 93.33%	• Combining style-based and content-based features is more accurate than traditional methods.
			• Functional word		• Acc_g= 83.52%	• The BERT model require more fine-tuning.
			• Special character			
			• Syllable			
	Hybrid-based	Garibay et al. [45]	• Punctuation	RF	• English:	• The proposed system presented some failures with the gender classification which affected the model performance.
			• Emoticons		Acc_g=70.6%	• The model suffers from overfitting problem caused by the number of estimators in the RF model.

Table 9 (continued)

Dataset	Approach	Reference	Features	Methods	Obtained results	Conclusion
			• Sentences		Acc_a=61.2%	
			• Numbers		RMSE=0.1749	
			• Links			
			• Words			
			• Letters			
	Hybrid-based	Basti et al. [10]	• Punctuation	• SVM	• Acc_a= 73.49%	• Stylometric features are more accurate than semantic and emotional features for user profiling.
			• Time of post	• MLP	• Acc_g= 83.70%	• The SVM classifier is the best classifier for user profiling.
			• Followers	• RF	• Acc_t = 88.70%	
			• Character			
			• Emotion			
			• Retweet			
			• Friends			
			• Topic			
			• Favorite account			
			• Words			
	Style-based	Basile et al. [9]	• Place names	SVM	Acc= 86%	• Place names are informative features for geographical location and language variety.
			• Characters			• Adding these features did not improve model performance.
			• Emojis			
			• Words			

Table 9 (continued)

Dataset	Approach	Reference	Features	Methods	Obtained results	Conclusion
	Content-based	Bayot et al. [11]	• Word2vec	SVM	• English:	• Using the Spanish Twitter dataset from PAN 2016, tf-idf worked better than word2vec in age classification.
			• TF-IDF		Acc= 70.2%	• Word2vec performed better for gender classification for both Spanish and English on the same dataset.
					• Spanish:	
					Acc= 56%	
	Hybrid-based	Gamallo et al. [25]	• References	NB	• English:	• Linguistic and lexical features are helpful for bot classification than for gender profiling.
			• URL links		Acc_b= 81%	• Combining lexical and textual features with BOW decreases the accuracy of the model.
			• Emoticons		Acc_g= 72%	
			• Retweets		• Spanish:	
			• Hashtags		Acc_b= 88%	
			• Content		Acc_g= 71%	
			• Emojis			
	Content-based	Sierra et al. [69]	• characters	CNN	• English:	• Using word sequences as input for the CNN was better than using sequences of characters.
			• Words		Acc=66%	• The CNN classifier produces additional challenges such as hyperparameter tuning and quick overfitting.

Table 9 (continued)

Dataset	Approach	Reference	Features	Methods	Obtained results	Conclusion
					• Spanish: Acc=73% • Portuguese: Acc=81% • Arabic: Acc=57%	
	Content-based	Kodiyan et al. [32]	• Words	RNN	• Acc_g= 75.31% • Acc_l= 85.22%	• The RNN exceeds the CNN in gender prediction when using the PAN 2017 training data.
	Content-based	Dias et al. [19]	• Characters	• CNN	• English: Acc_b=84% Acc_g=58% • Spanish: Acc_b=82% Acc_g=65%	• Gender profiling is more challenging than bot recognition task. • The Ensemble CNN-RNN-char-word model outperformed all alternatives.
			• Words	• RNN		
Social media	Content-based	Anjum et al. [5]	• Content	• SMO	• For PAN16:	• The classifier SMO performs well in gender identification task.
			• Words	• ID3	Acc_g= 49.8%	• The results of the PAN-AP16 are not much good.
				• J48 • NB • RF	Acc_a= 41.8%	

Table 9 (continued)

Dataset	Approach	Reference	Features	Methods	Obtained results	Conclusion
	Style-based	Mendoza et al. [42]	• Punctuation	SVM	• For Twitter:	• Personal phrases are very useful for predicting age and gender of social media users.
			• Slang words		Acc_a= 37.49%	• The plural personal phrases have not a special relevance for the AP task.
			• Stopwords		Acc_g= 59.55%	
			• Emoticons			
			• POS tags			
			• Words			
Facebook	Hybrid-based	Fatima et al. [21]	• Characters	• J48	• Acc_g= 87.5%	• Content based methods outperformed stylistic based methods for both age and gender classification tasks.
			• words	• RF	Acc_a= 75%	• Translation of multilingual corpus to monolingual text does not improve results.
			• Vocabulary richness	• SVM		
			• Lexical	• NB		
	Style-based	Rangel et al. [60]	• Punctuation	• SVM	• Acc_g= 59%	• Combining stylistic features with SEL dictionary gives competitive results.
			• Emoticons	• J48	• $F1_e$= 40.3%	• The manual labeling process was difficult and time consuming.
			• POS	• NB		
				• BN		

Table 9 (continued)

Dataset	Approach	Reference	Features	Methods	Obtained results	Conclusion
Blog	Hybrid-based	Mechti et al. [38]	• Prepositions	NB	Acc_a= 61.76 %	• The use of lexical classes is not enough for a good age classification.
			• Determiners			
			• Pronouns			
			• Adverbs			
			• Verbs			
Email	Hybrid-based	Safara et al. [64]	• Characters	• ANN	• Acc_g= 98%	• The proposed method outperformed traditional machine learning methods in terms of accuracy.
			• Structural	• SVM	• Pre_g= 97.16%	• The problem of this method was the execution time of ANN itself and then in combination with WOA.
			• Syntactic	• NB	• Rec_g= 99.67%	
			• Words	• DT		

Acc: Accuracy, Acc_g: Accuracy for gender classification, Acc_a: Accuracy for age classification, Acc_b: Accuracy for bot classification, Acc_e: Accuracy for emotions classification, Acc_l: Accuracy for language classification, Acc_o: Accuracy for occupation classification, Acc_f: Accuracy for fame classification, Acc_{bi}: Accuracy for birthyear classification, PC: Pearson Correlation, $F1_g$: F1-score for gender classification, $F1_b$: F1-score for bot classification, $F1_a$: F1-score for age classification, $F1_t$: F1-score for terrorist classification, Pre_g: Precision for gender classification, Rec_g: Recall for gender classification, RMSE: Root-Mean-Square Error

have so far focused on using monolingual text only. For example, all the PAN AP corpora are monolingual. But in social media, we noticed that most profile contents are written in multiple languages. To the best of our knowledge, there is no AP corpus available including profiles with multilingual texts. Therefore, there is a need to pay more attention to multilingual datasets in order to solve this problem. However, a recent work conducted by Mehwish et al. [21] focused on age and gender identification on multilingual Facebook corpus and did not provide a sufficient solution to solving the problem. Consequently, more work and research in multilingual AP are needed.

9 Conclusion and prospects

9.1 Summary

Due to the availability of huge unstructured data on social media networks, add to this the great importance and the need to carry out profiles identification tasks, a demand for methods and techniques capable of profiling users in these online platforms are constantly growing. In this paper, we have provided an overview of the process of author profiling on social networks. To provide a comprehensive overview of existing approaches, we proposed a taxonomy which focused on the type of the features used in each method. Thereafter, we presented the main techniques used for the classification of the authors. Machine learning and deep learning are the two mostly used techniques in the literature. Additionally, we analyzed and discussed the most relevant work studied in the literature to give researchers, in this field, a good comprehensive on the effective tools to solve the AP in social media. The synthesis assessment, carried out at the end of this work, has prompted us to introduce the main challenges encountered by researchers in this issue. We really hope that this article will provide a coherent understanding of this interesting research topic and be helpful for researchers to pursue future research in this domain.

9.2 Prospects

Based on the review of this study, there are still open challenges that need to be addressed. These challenges provide some open research directions that can motivate and help further researchers in advancing the AP task. Therefore, we propose below some promising orientations that could address these challenges.

Our suggestions for future research are structured around three directions. The first direction is to conduct an in-depth study of fake profile detection methods. Indeed, because of the rapid growth of social media networking, therefore due to the increase in the amount of personal information sharing among friends on these online platforms, protecting the privacy of individuals has become a serious challenge. Fake profiles constitute a very important issue in these collaborative environments. So, there is a need to develop an efficient and automatic method for the detection of fake profiles from different kinds of texts such as Twitter tweets, Facebook posts, LinkedIn comments, etc., and also differentiate the fake profiles from the authentic ones. As a second direction, we propose to find an effective mechanism for language independence that will be able to analyze users' profiles content in social networks in any language. Indeed, most social media content is displayed in multiple languages. Existing works on the analysis of multilingual texts did not provide efficacious and sufficient tools to solve this problem. Consequently, there is a need to shift attention to multilingual AP problems. The third direction that we suggest is to to adopt ontology-based

approaches to address the AP task. Recently, ontology has been used for the classification of scientific data in many different domains, but not yet for the AP problem.

Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

References

1. Abbasi A, Chen H (2008) Writeprints: a stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Trans Inf Syst (TOIS) 26(2):1–29
2. Akimushkin C, Amancio DR, Oliveira ON Jr (2018) On the role of words in the network structure of texts: application to authorship attribution. Phys A 495:49–58. https://doi.org/10.1016/j.physa.2017.12.054
3. Álvarez-Carmona MA, López-Monroy AP, Montes-y Gómez M, Villasenor-Pineda L, Meza I (2016) Evaluating topic-based representations for author profiling in social media. In: Ibero-American Conference on Artificial Intelligence. Springer, p 151–162
4. Alvarez-Carmona MA, Pellegrin L, Montes-y Gómez M, Sánchez-Vega F, Escalante HJ, López-Monroy AP, Villaseñor-Pineda L, Villatoro-Tello E (2018) A visual approach for age and gender identification on twitter. J Intell Fuzzy Syst 34(5):3133–3145. https://doi.org/10.3233/JIFS-169497
5. Anjum MW, Cheema WA (2018) A study of content based methods for author profiling in multiple genres. Int J Sci Eng Res 9:322–327
6. Ashraf S, Iqbal HR, Nawab RMA (2016) Cross-genre author profile prediction using stylometry-based approach. In: CLEF (Working Notes). Citeseer, p 992–999
7. 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)
8. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate
9. Basile A, Dwyer G, Medvedeva M, Rawee J, Haagsma H, Nissim M (2017) N-gram: new groningen author-profiling model. https://arxiv.org/abs/1707.03764
10. Basti R, Jamoussi S, Charfi A, Ben Hamadou A (2019) Arabic twitter user profiling: application to cyber-security, pp 110–117, https://doi.org/10.5220/0008167401100011
11. Bayot R, Gonçalves T (2016) Multilingual author profiling using word embedding averages and svms. In: 2016 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA). IEEE, p 382–386
12. Bentolila I, Zhou Y, Ismail LK, Humpleman R (2011) System, method, and software application for targeted advertising via behavioral model clustering, and preference programming based on behavioral model clusters. Google Patents. US Patent 8,046,797
13. Bilal M, Israr H, Shahid M, Khan A (2016) Sentiment classification of roman-urdu opinions using naïve bayesian, decision tree and knn classification techniques. J King Saud Univ-Comput Inf Sci 28(3):330–344
14. Bougiatiotis K, Krithara A (2016) Author profiling using complementary second order attributes and stylometric features. In: CLEF (Working Notes). p 836–845
15. Boukhari K, Omri MN et al Approximate matching-based unsupervised document indexing approach: application to biomedical domain
16. Bsir B, Zrigui M (2018) Enhancing deep learning gender identification with gated recurrent units architecture in social text. Computación Sistemas 22(3):757–766
17. Cui L, Zhang X, Qin AK, Sellis T, Wu L (2017) Cds: collaborative distant supervision for twitter account classification. Expert Syst Appl 83:94–103. https://doi.org/10.1016/j.eswa.2017.03.075
18. Daneshvar S, Inkpen D (2018) Gender identification in twitter using n-grams and lsa. In: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)
19. Dias RFS, Paraboni I (2019) Combined cnn+ rnn bot and gender profiling. In: Conference and labs of the evaluation forum (Working Notes)
20. Escalante HJ, Montes-y Gómez M, Villaseñor-Pineda L, Errecalde ML (2015) Early text classification: a naïve solution

21. Fatima M, Hasan K, Anwar S, Nawab RMA (2017) Multilingual author profiling on facebook. Inf Process Manag 53(4):886–904. https://doi.org/10.1016/j.ipm.2017.03.005
22. Fernquist J (2019) A four feature types approach for detecting bot and gender of twitter users. In: Working notes of CLEF 2019 - conference and labs of the evaluation forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of CEUR Workshop Proceedings. CEUR-WS.org
23. 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
24. Fourkioti O, Symeonidis S, Arampatzis A (2019) Language models and fusion for authorship attribution. Inf Process Manag 56(6):102061. https://doi.org/10.1016/j.ipm.2019.102061
25. Gamallo P, Almatarneh S (2019) Naive-bayesian classification for bot detection in twitter. In: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9–12, 2019, volume 2380 of CEUR Workshop Proceedings. CEUR-WS.org
26. Giachanou A, Ríssola EA, Ghanem B, Crestani F, Rosso P (2020) The role of personality and linguistic patterns in discriminating between fake news spreaders and fact checkers. In: International Conference on Applications of Natural Language to Information Systems. Springer, p 181–192
27. Johansson F (2019) Supervised classification of twitter accounts based on textual content of tweets. In: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9–12, 2019, volume 2380 of CEUR Workshop Proceedings. CEUR-WS.org
28. Joo Y, Hwang I (2019) Author profiling on social media: an ensemble learning model using various features, 2380
29. Juola P (2015) Industrial uses for authorship analysis. Math Comput Sci Ind 1:21–25
30. Kaati L, Lundeqvist E, Shrestha A, Svensson M (2017) Author profiling in the wild. In: 2017 European Intelligence and Security Informatics Conference (EISIC). IEEE, p 155–158
31. Kapociute-Dzikicne J, Damaševicius R (2018) Lithuanian author profiling with the deep learning. In: 2018 Federated Conference on Computer Science and Information Systems (FedCSIS 2018), pp 169–172
32. Kodiyan D, Hardegger F, Neuhaus S, Cieliebak M (2017) Author profiling with bidirectional rnns using attention with grus: Notebook for pan at clef 2017. In: CLEF 2017 Conference and Labs of the Evaluation Forum, Dublin, Ireland, 11-14 September 2017, vol 1866. RWTH Aachen
33. Kovács G, Balogh V, Mehta P, Shridhar K, Alonso P, Liwicki M (2019) Author profiling using semantic and syntactic features: Notebook for pan at clef 2019, 2380
34. Kudugunta S, Ferrara E (2018) Deep neural networks for bot detection. Inf Sci 467:312–322
35. Lakkaraju SK, Tech D, Deng S (2018) A framework for profiling prospective students in higher education. In: Encyclopedia of Information Science and Technology, Fourth Edition. IGI Global, p 3861–3869
36. Mabrouk O, Hlaoua L, Omri MN (2018) Fuzzy twin svm based-profile categorization approach. In: 2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). IEEE, p 547–553
37. Mabrouk O, Hlaoua L, Omri MN (2018) Profile categorization system based on features reduction. In: International Symposium on Artificial Intelligence and Mathematics, ISAIM 2018, Fort Lauderdale, Florida, USA, January 3–5, 2018
38. Mechti S, Jaoua M, Faiz R, Bouhamed H, Belguith LH (2016) Author profiling: age prediction based on advanced bayesian networks. Res Comput Sci 110:129–137
39. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space
40. Moreno-Sandoval LG, Puertas E, Plaza-Del-Arco FM, Pomares-Quimbaya A, Alvarado-Valencia JA, Ureña-López A (2019) Celebrity profiling on twitter using sociolinguistic features notebook for pan at clef 2019
41. 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, volume 1391 of CEUR Workshop Proceedings. CEUR-WS.org
42. Ortega-Mendoza RM, Franco-Arcega A, López-Monroy AP, Montes-y Gómez M (2016) I, me, mine: the role of personal phrases in author profiling. In: International Conference of the Cross-Language Evaluation Forum for European Languages. Springer, p 110–122
43. Ortega-Mendoza RM, López-Monroy AP, Franco-Arcega A, Montes-y Gómez M (2018) Emphasizing personal information for author profiling: new approaches for term selection and weighting. Knowl-Based Syst 145:169–181. https://doi.org/10.1016/j.knosys.2018.01.014
44. Ouni S, Fkih F, Omri MN (2021) Toward a new approach to author profiling based on the extraction of statistical features. Soc Netw Anal Min 11(1):1–16

45. Palomino-Garibay A, Camacho-González AT, Fierro-Villaneda RA, Hernández-Farias I, Buscaldi D, Meza-Ruiz IV (2015) A random forest approach for authorship profiling?notebook for pan at clef 2015. Work Notes Pap CLEF, 1391
46. Para U, Patel MS (2021) A new feature selection technique for author profiling. Des Eng 6:2868–2885
47. Park G, Schwartz HA, Eichstaedt JC, Kern ML, Kosinski M, Stillwell DJ, Ungar LH, Seligman MEP (2015) Automatic personality assessment through social media language. J Pers Soc Psychol 108(6):934
48. Pennacchiotti M, Popescu A-M (2011) Democrats, republicans and starbucks afficionados: user classification in twitter. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, San Diego, pp 430–438
49. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), p 1532–1543
50. Posadas-Durán J-P, Markov I, Gómez-Adorno H, Sidorov G, Batyrshin I, Gelbukh A, Pichardo-Lagunas O (2015) Syntactic n-grams as features for the author profiling task. Work Notes Pap CLEF, 1391
51. Poulston A, Waseem Z, Stevenson M (2017) Using tf-idf n-gram and word embedding cluster ensembles for author profiling. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11–14, 2017, volume 1866 of CEUR Workshop Proceedings
52. Prasad SN, Narsimha VB, Reddy PV, Babu AV (2015) Influence of lexical, syntactic and structural features and their combination on authorship attribution for telugu text. Procedia Comput Sci 48:58–64. https://doi.org/10.1016/j.procs.2015.04.110
53. Puertas E, Moreno-Sandoval LG, Plaza-Del-Arco FM, Alvarado-Valencia JA, Pomares-Quimbaya A, Ureña-López A (2019) Bots and gender profiling on twitter using sociolinguistic features notebook for pan at clef 2019, 2380
54. 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. https://www.sciencedirect.com/science/article/abs/pii/S0306457315000783
55. Rangel F, 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, volume 2380 of CEUR Workshop Proceedings, pp 1–36. CEUR-WS.org
56. Rangel F, Rosso P, Charfi A, Zaghouani W, Ghanem B, Snchez-Junquera J (2019) Overview of the track on author profiling and deception detection in arabic. In: Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. CEUR-WS.org, Kolkata, India
57. Rangel F, Rosso P, Chugur I, Potthast M, Trenkmann M, Stein B, Verhoeven B, Daelemans W (2014) Overview of the 2nd author profiling task at pan 2014. In: CLEF 2014 Evaluation labs and workshop working notes papers. Sheffield, pp 1–30
58. 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, p 352–365
59. Rangel F, Rosso P, Potthast M, Stein B (2017) Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. Work Notes Pap CLEF 48:1613–0073
60. Rangel Pardo F, Rosso P (2013) On the identification of emotions and authors' gender in facebook comments on the basis of their writing style. CEUR Work Proc CEUR-WS 1096:34–46
61. Rangel Pardo FM, 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, volume 1391 of CEUR Workshop Proceedings, pp 1–8. CEUR-WS.org
62. Rico-Sulayes A (2011) Statistical authorship attribution of mexican drug traficking online forum posts. Int J Speech Lang Law 18(1):53–74
63. Rosso P, Rangel F (2020) Author profiling tracks at fire. SN Comput Scie 1(2):1–11. https://link.springer.com/article/10.1007/s42979-020-0073-1
64. 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
65. Sboev A, Litvinova T, Gudovskikh D, Rybka R, Moloshnikov I (2016) Machine learning models of text categorization by author gender using topic-independent features. Procedia Comput Sci 101:135–142
66. Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, Shah A, Kosinski M, Stillwell D, Seligman MEP et al (2013) Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS ONE 8(9):e73791

67. 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 Humaniz Comput 10(9):3409–3425. https://doi.org/10.1007/s12652-018-1050-6
68. Sharjeel M, Fatima M, Anwar S, Nawab RMA (2018) Multilingual author profiling on sms track at fire'18. In: Proceedings of the 10th annual meeting of the Forum for Information Retrieval Evaluation, FIRE 2018, Gandhinagar, India, December 06-09, 2018, pp 16–17
69. Sierra S, Montes-y Gómez M, Solorio T, González FA (2017) Convolutional neural networks for author profiling. Work Notes CLEF
70. Soler J, Wanner L (2016) A semi-supervised approach for gender identification. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23–28, 2016, pp 1282–1287. European Language Resources Association (ELRA)
71. Takahashi T, Tahara T, Nagatani K, Miura Y, Taniguchi T, Ohkuma T (2018) Text and image synergy with feature cross technique for gender identification
72. Villena-Román J, Cristóbal JCG (2014) Daedalus at pan 2014: guessing tweet author's gender and age. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15–18, 2014, volume 1180 of CEUR Workshop Proceedings, pp 1157–1163. CEUR-WS.org
73. Yang M, Chen X, Tu W, Lu Z, Zhu J, Qu Q (2018) A topic drift model for authorship attribution. Neurocomputing 273:133–140. https://doi.org/10.1016/j.neucom.2017.08.022
74. Zhang W, Caines A, Alikaniotis D, Buttery P (2016) Predicting author age from weibo microblog posts. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). p 2990–2997