Expert Systems With Applications 199 (2022) 117140

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Survey on profiling age and gender of text authors

Yaakov HaCohen-Kerner

Department of Computer Science, Jerusalem College of Technology – Lev Academic Center, 21 Havaad Haleumi St., POB 16031, 9116001 Jerusalem, Israel

ARTICLE INFO

Keywords:
Age classification
Author profiling
Deep learning
Gender classification
Supervised machine learning
Text classification

ABSTRACT

Author profiling from text documents has become a popular task in latest years, in natural language applications. Author profiling is important for various domains such as advertising, marketing, forensics, and security. This survey focuses on profiling age and gender, the two features, which are probably the most researched profile attributes. In this paper, we present an overview of representative studies and datasets of the field (including those organized by PAN) with several significant leaps. Due to the increasing use of deep learning (DL) methods in recent years, we have also reviewed several DL systems that profile authors' age and gender. Most age and gender datasets contain blog posts or Twitter messages written in English, Spanish or Arabic. There are also several relevant datasets written in Dutch, Italian, Portuguese, Turkish, and Russian. There is no consistency and no uniformity in the datasets concerning to the number and types of their documents, the division into training, dev, and test sets, the types of the applied preprocessing methods, and the quality measures used to evaluate the classification results. A prominent interesting finding is that the best age accuracy results are not as high as we might have expected taking into account relatively simple types of classification especially by gender (only 2 categories) when a large number of teams have competed over the years. Another interesting finding that repeats itself in various classification tasks is that classical ML methods are still better than DL methods for age and gender classification tasks. Most classical systems used word unigrams and bigrams and character 3–4-5-grams. Several systems also used various types of stylistic features. While many earlier systems did not apply pre-processing methods, most recent systems applied several preprocessing methods, e.g., lowercase conversion and replacement of various strings (e.g., URLs, LF characters, and User Mentions). We also suggest several potential future issues in age and gender profiling research.

1. Introduction

Groups of people with different demographics differ systematically in their behavior, including their language behavior. Sociolinguists have accumulated a large body of evidence about language variation correlated with sociological dimensions such as age, gender, and educational background (e.g., Coupland, 2007). Similarly, language psychologists have found correlations between language use and psychological traits (Pennebaker, 2011). Author profiling (AP, also called authorship characterization), can be defined as predicting the attributes of the author of a textual document, where each attribute can be either biological or socio-cultural. In automatic profiling of authors from text, the goal is to start from the language behavior observed, analyze its linguistic properties, and determine the profile that fits that type of language use.

Various author characteristics have been explored in textual datasets during the last two decades, e.g., age (Koppel et al., 2006; Argamon et al., 2007; Lpez-Santamara et al., 2019), educational background, (Coupland, 2007), educational level (Juola and Baayen, 2003), ethnic origin (HaCohen-Kerner et al., 2008A; HaCohen-Kerner et al., 2010A; HaCohen-Kerner et al., 2010B), gender (Koppel et al., 2002; Argamon et al., 2003; Koppel et al., 2006; Kucukyilmaz et al., 2006; Argamon et al., 2007; Cheng et al., 2009; Cheng et al., 2011; Lpez-Santamara et al., 2019), language background (de Vel et al., 2002; Zheng et al., 2006), location (Reddy et al., 2016), native language (Koppel et al., 2005), neuroticism level (Argamon et al., 2009), and personality type (Pennebaker et al., 2003; Noecker et al., 2013; Verhoeven et al., 2016).

A domain related to AP is stylometry, which was defined by Zheng et al. (2006) as the statistical analysis of writing style and by Nerbonne (2014) as using general techniques to explore textual style. About one decade earlier, Holmes (1994; 1998) defined features that quantify writing style, a research approach known as "stylometry". Many types of measures have been proposed, e.g., character frequency, sentence length, vocabulary richness functions, word frequency, and word length.

There is currently a large interest in reliable approaches to AP, a

E-mail address: kerner@jct.ac.il.

https://doi.org/10.1016/j.eswa.2022.117140
Received 10 June 2021; Received in revised form 19 January 2022; Accepted 29 March 2022
Available online 5 April 2022

subject with many commercially and societally relevant applications. In business intelligence, these approaches can be used to analyze demographically differentiated opinions about products and companies on social media, and in targeted marketing and advertising. They can also be used in customer relations to guess the personality of patrons that use a company's conversational agent by adapting the conversational style of the chatbot to this personality profile. In forensics, profiles provided online can be checked based on language use, and the author of letters, e-mails, and other documents that are part of an investigation can be profiled.

The increase of interest in research in AP can be seen in Figs. 1-3, where we track the number of publications in Google Scholar having "author profiling," "age and classification/categorization," and "gender and classification/categorization," respectively, in their title, in the years 2007–2020[1].

Analysis of Figures 1–3.

Fig. 1 presents the number of publications in Google Scholar having "author profiling and classification/categorization" in their title over the years 2007–2020. In 2007–2012, the number of such articles was pretty low (<=4). In 2013, the number of such publications has been dramatically increased. The main reason for this is probably the establishment of the first AP task at PAN[2] in 2013 (Rangel et al., 2013). PAN is a series of scientific events and shared tasks on digital text forensics and stylometry. Since 2010, the CLEF Initiative (Conference and Labs of the Evaluation Forum) has hosted PAN evaluation. In 2014–2015, there was a decrease in the number of such articles. In 2016–2018, there has been a moderate increase. In 2019–2020, there was another decrease. In our experience, the results for 2019–2020 are unstable and do not reflect the truth; these results are expected to increase.

On the other hand, Figs. 2-3, which present the number of publications in Google Scholar having "age and classification/categorization" and "gender and classification/categorization" in their title over the years 2007–2020, respectively, introduce we find a fairly stable gradual rise in the number of articles. As mentioned before, according to our experience, the results for 2019–2020 are still unstable and likely to increase. A possible reason for the differences between Figs. 2-3 and Fig. 1 is that age and gender classification tasks were not "born" in 2013 but were "alive" at least since 2007 and therefore the gradual increase in their number over the years is quite stable. In contrast, the concept of author profiling classification was "born" in 2013 because of the organization of the first AP task at PAN in 2013.

Most current approaches to AP use supervised machine learning (ML) methods. Textual datasets are collected with associated profile information. An example of a suitable dataset is a blog dataset that provides texts and profile information such as age, gender, and location. The texts are provided automatically with a linguistic analysis (feature construction), to make possible the construction of feature vectors. Feature types in general, and specific features in particular, range from superficial (average sentence length, average word length, character count, punctuation mark count, word count, and vocabulary richness features) to complex (e.g. discourse structure and propositional density). This data, feature vectors as input with associated profile dimensions as output, is used by supervised ML methods to construct and evaluate models that can classify the profile of previously unseen texts. Research in this area is focused on (i) searching for good combination(s) of predictive features, (ii) selecting appropriate supervised ML methods for solving specific profiling tasks with maximal accuracy, and (iii) finding explanations for why specific models perform well.

In this paper, we present, an overview of age and gender profiling from text. We decided to concentrate on these two features/characteristics because (1) they are probably the two most popular features/characteristics (or at least among the most popular ones) in AP studies; (2) these two profile dimensions are well researched, and in which robust results have been achieved for many text datasets in various languages and domains; and (3) these profile dimensions are important for various types of applications, e.g., AP (Fatima et al., 2017; Koch et al., 2020), adversarial stylometry (Brennan et al., 2012; Emmery et al., 2021), bias elimination (Chopra et al., 2020), characterization of the differences between types of gender and age (Kumar et al., 2020; Foong and Gerber, 2021), detection of sexual harassment (Karami et al., 2020), and targeted advertising (Stefanija, 2021).

In this paper, we summarize or present information from various survey/review papers as well as many research papers. These papers are usually dedicated to specific tasks, languages, datasets, competitions, and systems and are updated until the articles are published (in 2019 or earlier). In contrast, we do not limit ourselves to any specific task, language, dataset, competition, ML method, and system. We try to do our best to present a comprehensive survey paper updated until 6/Jun/2021, regarding gender and/or age classification tasks without any kind of restriction.

In Section 2, we provide an overview of previous overview papers and in Section 3, we provide an overview of previous papers related to gender and/or age identification/classification. Section 4 presents previous representative datasets is presented. Section 5 summarizes PAN's competitions and winners. Section 6 presents several systems that profile authors' age and gender using deep learning methods. Finally, in Section 7, we summarize the findings, conclude, and offer several ideas for further research.

2. Overview of previous overview papers

Different research domains such as computational linguistics, linguistics, and psychology try to discover which linguistic features characterize the profile of either text authors or speakers. Pennebaker et al. (2003) summarized some of the linguistic features that characterize the personality of text authors. Specifically, the authors indicated the significance of articles, auxiliary verbs, conjunctions, particles, prepositions, and pronouns. In the English language, for instance, although there are less than 200 commonly used particles, they constitute over half of the words used. The authors claim in their summary, based on correlations between language use and gender, that women are inclined to use more personal pronouns and words that relate to emotions, while men are inclined to use more articles, long words (words containing more than six characters), nouns, and prepositions (because they are inclined to describe their environment in greater depth). Regarding age, Pennebaker et al. (2003) discovered that younger authors use more first-person pronouns and verbs in the past tense, while older authors are inclined to use more articles, nouns, prepositions, and verbs in the future tense. Below, we show previous overview papers and their limitations. Rosso et al. (2018) reviewed the state of the art of several AP tasks (as well as some other tasks), especially in Arabic. The authors presented summaries of four studies dealing with age and gender identification in Arabic between 2008 and 2016. The examined datasets were: a few thousand emails, a few hundreds of articles collected from well-known Arabic newsletters, and a few thousands of tweets. For each study, the authors presented its extracted features, applied supervised ML methods, and accuracy results. The limitations of this study are: (1) it deals mainly with studies in Arabic; (2) the review about age and gender in this study is quite limited (about 3 pages; one page for PAN shared tasks and two pages about related Arabic studies); (3) there are no reviews about systems that applied deep learning and/or word embedding methods; and (4) it deals with studies between 2008 and 2016.

Eke et al. (2019) presented a survey of state-of-the-art approaches for user profiling mainly from the viewpoints of user profile types (static and dynamic), user profile models (behavioral, interest, and intention), and user profiling process (user profile construction, user information collection methods (implicit and explicit methods), and user profile updating methods. The authors summarized the modeling process from

[1] Figures 1–3 are updated to 6-Jun-2021.

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

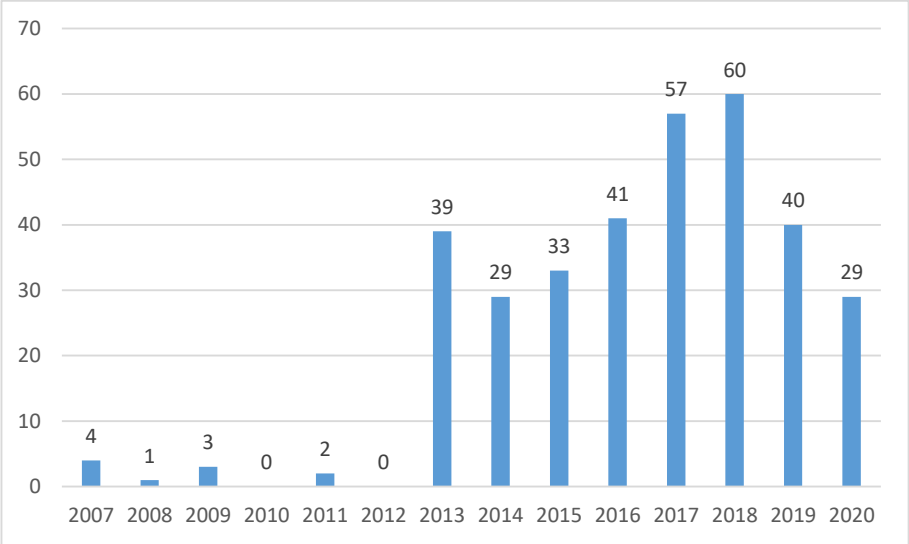

Fig. 1. The number of publications in Google Scholar having "author profiling" in their title in the years 2007–2020.

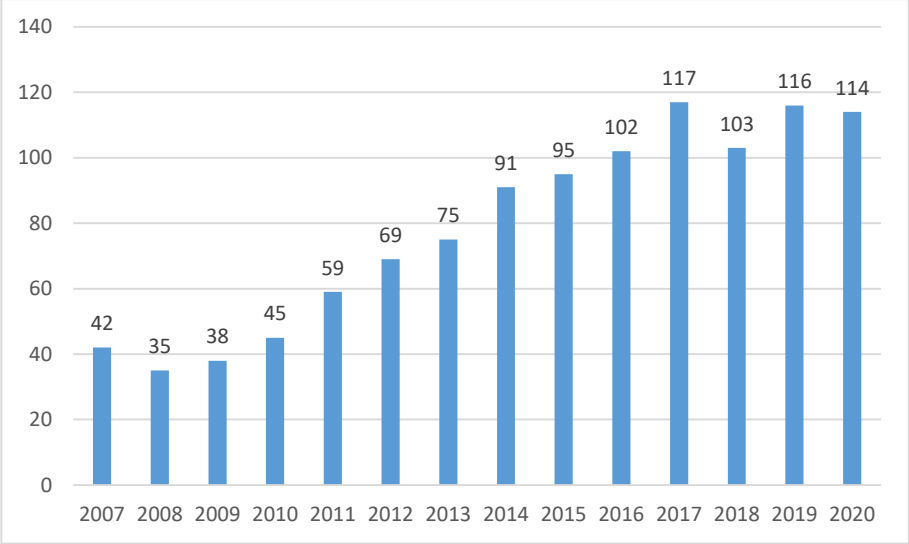

Fig. 2. The number of publications in Google Scholar having "age and classification/categorization" in their title in the years 2007–2020.

the viewpoints of the data sources, extracted features, profiling approaches, and evaluation metrics. In addition, the authors presented the weaknesses and strengths of each approach and discussed the open challenges. The limitations of this study are: (1) it concentrates on approaches for user profiling mainly from the viewpoints of user profile types, i.e., there are no details about age and gender studies; (2) there are no reviews about systems that applied deep learning and/or word embedding methods and (4) it deals with studies until 2018 (only one reference from 2019).

Rangel et al. (2019) presented the AP and deception detection for the Arabic shared task at PAN@FIRE 2019. A balanced ARAP-Tweet dataset was built for the AP task. It contains 15 dialects related to 22 Arab countries. For each dialect, a total of 198 authors with more than 2,000 tweets for each author, were annotated with gender and age (under 25, between 25 and 34, and over 35). 28 teams participated in the AP task. 15 notebook papers were submitted and accepted. The published studies were analyzed according to the following viewpoints: preprocessing methods, used features, applied ML methods, and accuracy results. The most popular preprocessing method was the removal of Arabic stop words. Several preprocessing methods were applied by two or three teams, e.g., removal of special characters, punctuation signs, numbers, non-Arabic words, and various Twitter items e.g., emojis, hashtags, URLs, and user mentions. Only one team applied tokenization and only one team lowercased the texts. The participated teams used different types of features, e.g., character/word n-grams, stylistic features; and embeddings. Most teams used traditional ML methods, e.g., support vector machine (SVM), random forest (RF), logistic regression (LR), and Multinomial Naive Bayes (MNB) (Kibriya et al., 2004). Only two systems used deep learning (DL) methods (Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) and long short term memory (LSTM) (Sherstinsky, 2020). The traditional ML methods achieved better accuracy results than the DL ones. The best performing team for the gender task obtained an accuracy of 0.8194 by applying RF using a combination of words and emoticons 2-grams and 3-grams. The best performing team for the age task obtained an accuracy of 0.6250 by applying LR using a combination of word unigrams and character 2 to 5-grams. The limitations of this study are: (1) it deals with studies in Arabic only; (2) it deals with a dataset of tweets; and (3) it deals with studies from 2019 only.

In contrast to these previous overview papers, in this study, we

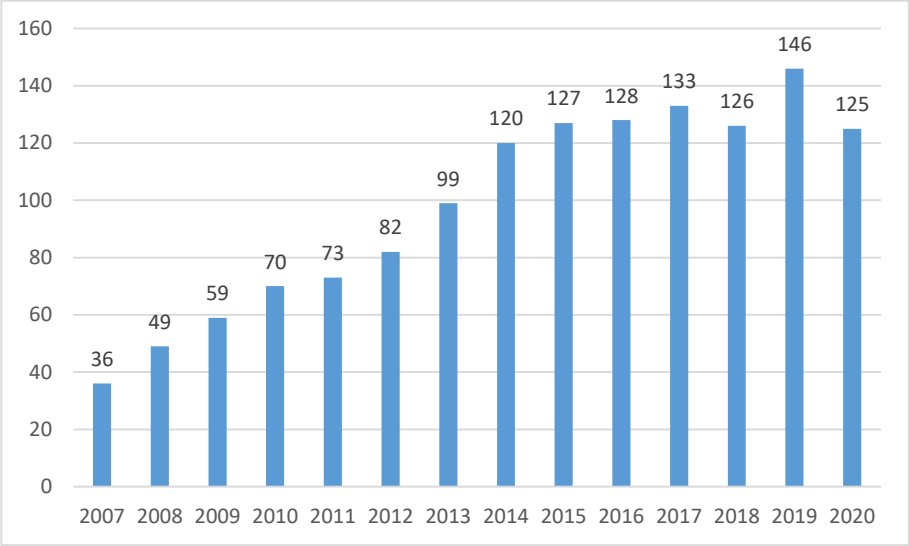

Fig. 3. The number of publications in Google Scholar having "gender and classification/categorization" in their title in the years 2007–2020.

present an overview of representative studies and datasets of the field without any restriction regarding specific age and/or gender task, language, dataset, year, competition, ML method, and systems. In this review, we will try to point out interesting characteristics and findings of many studies in the field.

3. Overview of previous papers related to gender and/or age Identification/Classification

Koppel et al. (2002) performed classification experiments on a partial dataset of the British National Corpus (BNC) corpus containing 566 documents: 264 fiction documents and 302 non-fiction documents with an equal number of documents authored by males and females for each type of document, achieving an accuracy of 77.3% using the Exponential Gradient (EG) ML method with 1,081 features: 405 function words, and 676 PoS-Tags n-grams: 500 triples, 100 pairs, and 76 single tags.

Argamon et al. (2003) analyzed 604 BNC documents: 246 fiction documents and 358 non-fiction documents with an equal number of documents authored by males and females for each type of document, achieving an accuracy of about 80% in gender classification. The authors discovered notable differences between documents authored by males and females in the use of specific types of noun modifiers and pronouns. It was found that even in formal writing, female writing displays a wider usage of features that researchers identified as "involved," while male writing displays a greater usage of features identified as "informational." For instance, they discovered that males make much wider use of noun specifiers, while females make much wider use of pronouns.

Schler et al. (2006) constructed a sub-corpus from Blogger.com in Aug 2004, which includes more than 71,000 blogs, containing 681,288 blog posts with over 140 million words. These researchers downloaded these blog posts. The gender of the bloggers was not balanced: 37,324 are males and 34,169 are females. Females are a distinct majority (63%) of bloggers up to the age of 17, but a minority (47%) of bloggers above the age of 17. To balance, they created a sub-corpus, which consisted of an equal number of male and female bloggers in each age group by randomly discarding surplus documents. This sub-corpus contains 37,478 blogs, 1,405,209 blog posts with almost 300 with over 140 million words.

For their classification tasks on this sub-corpus, they used 502 stylistic features such as blog words, part of speech (PoS) tags, hyperlinks, and 1,000 word unigrams, which yielded the greatest information gain in the training set. The results for gender identification, making use of style-related features, are approximately consistent with those achieved in previous researches on much smaller corpora of non-fiction and fiction documents (Koppel et al., 2002; Argamon et al., 2003). The authors note that although the prominent differences in content between female and male bloggers noted above, stylistic differences were more significant than content differences. Application of the multi-class balanced real-valued Winnow (MCRW) ML method using all the features obtained an accuracy of 80.1%. Age experiments were conducted on three age categories: teens (13–17), 20′s (23–27), and 30′s (33–42). They made use of these categories – omitting intermediate ages – to allow a clear differentiation between the categories, particularly because many of the blogs used had been active for several years. The baseline was simply predicting the majority class (teens), yielding an accuracy of 43.8%. Content features proved to be a bit more useful than stylistic features, but – as was the case in gender identification – the combination of both types of features was most useful. The confusion matrix indicates that using both types of features, teens can be distinguished from 30′s with an accuracy of 95.76%, and teens are distinguishable from 20′s with an accuracy of 87.3%. However, many 30′s were incorrectly classed as 20′s, resulting in an overall accuracy of 76.2%. The authors also found that within each age group, male and female bloggers use different content words and different blogging styles, e.g., regardless of gender, writing style grows increasingly "male" with age: pronouns and assent/negation become scarcer, while prepositions and determiners become more frequent.

Argamon et al. (2007) constructed from Blogger.com in Aug 2004 a sub-corpus of 681,288 blog posts containing more than 140 million words. These posts were collected from 19,320 blogs divided into 8,240 authored by 10′s (13–17), 8,086 authored by 20′s (23–27), and 2,994 authored by 30′s (33–42)). On average, each blog contained around 35 posts that contain 7,300 words. In their classification tasks, they used the 1,000 most frequent words: 323 function words and 677 content words. They applied two ML methods: Bayesian multinomial and logistic regression (BMR) (Madigan, et al., 2005) and MCRW (Littlestone, 1988). The best accuracy results were 77.4% for the age task using BMR, and 80.5% for the gender task using WIN.

The age-experiment results show significant differences in various content and style-related words between blog posts written by different-aged bloggers. The use of content words that are associated with *Business, Family, Internet, Politics,* and *Religion* increases with the increase in age, while the use of words associated with *AtHome, Conversation, Fun, Music, Romance, School,* and *Swearing* significantly decreases with the increase in age. Stylistic features measured by frequencies of

grammatical (PoS) tags show that the usage of auxiliary verbs, conjunctions, and personal pronouns significantly decreases with age, while the use of prepositions and articles significantly increases with the increase in age. The gender-experiment results also show significant differences in various content words and stylistic features. The usage of content words that are associated with *Business, Internet, Politics, and Religion,* are used more frequently by males, while words that are associated with *At Home, Conversation, Fun, Romance,* and *Swearing* are more used by females. Stylistic features *such as prepositions* and *articles* are significantly more used by males, while *auxiliary verbs, conjunctions,* and *personal pronouns* are significantly more used by females. These features are the same features that were found to indicate male and female writing styles in published fiction and non–fiction works (Argamon, et al., 2003). The authors also found that stylistic features and content words are highly correlated. Older bloggers and male bloggers use more *words* related to *Articles, Business, Internet, Politics, Prepositions,* and *Religion*, while younger bloggers and female bloggers, use more words related to *AtHome, Auxiliary Verbs, Conjunctions, Conversation, Fun, Personal Pronouns, Romance,* and *Swearing*. The Pearson correlation between the *male/female* and *30+/10 s* log–ratios 0.71 is considered relatively high.

Argamon et al. (2009) studied the task of automatic AP of an anonymous text from a few viewpoints: age, gender, native language, and personality. They used novel linguistic-based features that include both function words and PoS tags. They represent systemic functional linguistics (Halliday et al., 2014) as trees, and they label the leaves of these trees with sets of PoS tags. For each node in these trees, they compute the frequency of the word normalized by the number of words in the text. They also consider 1,000 content-based features (i.e. words that appear above a certain threshold and distinguish between the relevant categories.

They applied only one ML method; the Bayesian Multinomial Regression (BMR) (Genkin et al., 2007). They investigated four profiling tasks: age, gender, native language, and neuroticism level. They used three different corpora. They applied the age and gender experiments on the blog sub-corpus that was presented in (Schler et al., 2006) and applied the native language and neuroticism level experiments to the two other corpora. They ran 10-fold cross-validation tests for each of the following feature sets – content features only, stylistic features only, and both. For each category within each of the four tasks, they also presented the features, which were the most discriminating ones. The experimental results are as follows. Age: they labeled each blog in their corpus, based on the reported age of each blogger, as belonging to one of three age groups: 13–17 (42.7%), 23–27 (41.9%), and 33–47 (15.5%). Both content and style features obtain over 76.1% accuracy for this classification task, while the baseline majority-class classifier gave 42.7%. Contractions without apostrophes (younger writers), and determiners and prepositions (older writers) are the stylistic features that proved to be most useful for classification. Words related to mood and school for teens, to social life and work for writers in their twenties, and to family life for writers in their thirties were the content features that proved to be most useful for classification. Gender: prepositions and determiners (markers of male writing) and pronouns (markers of female writing) were the stylistic features that proved to be most useful for gender classification. Words related to technology (male) and words related to relationships and personal life (female) were the features that proved to be most useful for gender classification.

Various new features, such as average length of sentences, non-dictionary words that had an occurrence of greater than 50, and slang words, were added by Goswami et al. (2009). Working on the blog sub-corpus presented in (Schler et al., 2006), the authors applying the Naïve Bayes (NB) classifier, obtained an accuracy of 80% in age group detection by using 35 content words, 52 slang words, and average sentence length, and 89% in gender detection by using 35 content words and 52 slang words.

Burger et al. (2011) suggested a few versions of a language-independent classifier to classify the gender of users of Twitter. The examined dataset contains 4,102,434 tweets authored by 183,729 Twitter users (about 22 tweets per user) labeled with gender. This dataset was divided into three sub-datasets: (1) Training with 3,280,532 tweets authored by 146,925 Twitter users; (2) Development with 403,830 tweets authored by 18,380 Twitter users; and (3) Test with 418,072 tweets authored by 18,424 Twitter users. They used 15,572,522 features composed of character- and word-level n-grams generated from the following four information fields: full name, screen name, description, and tweet text. They applied only the balanced Winnow ML method. The best classifier using all four fields obtained an accuracy of 92% on the test sub-dataset, while the classifier that relied only on tweet texts obtained an accuracy of 75.5%. A large-scale human evaluation, which made use of Amazon Mechanical Turk, was performed. Only 6 people out of 130 people who performed at least 100 classification tasks achieved higher accuracy results than the best ML classifier. That is to say, their gender classification method significantly outperforms about 95% of humans.

Peersman et al. (2011) applied TC to predict gender and age on a dataset of chat texts, which were gathered from *Netlog*, the Belgian social networking site. This dataset contains 1,537,283 Flemish-Dutch posts. They improved on the random baseline performance for age classification making use of very limited datasets of 12.2 tokens (i.e. words, punctuation marks, and emoticons) per instance on average. Word unigrams such as "bro" (*brother*)) appear to be more helpful for age prediction than combing them in 2-grams or 3-grams. In the course of training on the 50,000 most informative word unigrams (χ^2), SVM displayed relatively high results for distinguishing between adults and adolescents. On a dataset of 10,000 posts per class, SVM achieved an accuracy of 71.3% for the task of discrimination between age classes of *min16* (from 11 to 15 years old) *and plus16* (16 years old and older). For more distant age groups, *min16* vs. *plus25,* the accuracy rose to 88.2 %. They used gender information for the age experiments. They subsequently presented their results of four experiments: comparing simultaneously *min16_male*, *min16_female*, *plus25_male*, *and plus25_female,* where each class contains 5,000 instances. Balancing their dataset according to both age and gender achieved the best results for the adult class, (accuracy − 88.8%, recall − 92.9%, precision − 91.5%, and F1-score − 91.7%). To gain greater insight into the minimum dataset size that would be needed in retraining experiments in the future, the authors also conducted experiments with different dataset sizes. When the experiment was conducted on a dataset half the size of the first dataset, SVM still achieved an accuracy of 87.7 % for distinguishing between *min16* and. *plus25*, indicating a very small decrease of 0.5% compared to the accuracy achieved on the initial dataset.

Cheng et al. (2009) researched gender detection for content-free, short, and multi-genre e-mails. For their research, they worked with the *Enron* e-mail dataset, which consists of 8,970 e-mails. The authors performed various experiments using 545 features (gender-linked, psycho-linguistic, and stylometric).). They found that both function words and word-based features play are significant in gender detection. The best accuracy results (82.2%)were obtained by SVM, which outperformed the decision tree method. The results of their experiments also point out that increasing the number of words in each e-mail, as well as the number of e-mails in the training dataset, improves the detection performance.

Mukherjee and Liu (2010) suggested two new methods to improve the gender classification of blog authors. One method presents variable-length PoS sequence patterns mined from the training set. These new features are capable to capture complex stylistic characteristics of female and male authors. The other method is a new feature selection method based on an ensemble of a few feature selection criteria and methods. Empirical evaluations using a real-life blog dataset, which consists of 3,100 blogs show that the classification accuracy results using these two methods are significantly higher than the results of the current state-of-the-art methods (e.g., Schler et al., 2006; Yan and Yan, 2006;

and Argamon et al., 2007). An accuracy result of 88.56% was obtained by the SVM regression ML method and around 20,000 PoS patterns.

Cheng et al. (2011) researched gender detection for content-free, short, and multi-genre text. Similar to the study of Cheng et al. (2009), the authors performed various experiments using 545 features (gender-linked, psycho-linguistic, and stylometric). In this study, the experiments were performed on two datasets: the newsgroup data *set in the Reuters Corpus Volume 1* (RCV1), which consists of 6,769 stories and the *Enron* e-mail dataset, which consists of 8,970 e-mails. The authors pointed out that SVM yields greater accuracy than the AdaBoost decision tree and the Bayesian-based logistic regression, with an accuracy result of 85.1% in gender detection. The experimental results also pointed out that function words, structural features, and word-based features, are important in gender detection. Cheng et al. (2011) also found that personally written texts such as e-mails and news articles show significant differences between females and mails. The experiments also show that increasing the number of words in each e-mail, as well as the number of e-mails in the training dataset, improves the detection performance.

Rangel and Rosso (2016) researched the influence of emotions on author gender and age identification. They proposed an emotion-based graph method that models the way people write to express themselves. They applied their method for gender and age identification on the Spanish sub-dataset of the PAN-AP-13 dataset, obtaining accuracy results comparable to those of the highest-performing systems on this sub-dataset. They hypothesized that males use more prepositional syntagmas than females, write on various topics using different emotions, and this will award a special significance to the sequence preposition + determinant + noun + adjective. Thus, they built a graph of different PoS tags of the texts of users and augmented it with semantic information: the topics they discuss, the type of verbs they make use of, and the emotions they convey. The whole text was modeled as only one single graph, taking into consideration punctuation marks as well, to denote how an author may begin and end sentences (i.e., how they link the concepts in their sentences). They extracted several properties from the generated graph and used them as features for an ML method. They found that females use more emotional verbs (e.g. want, feel, like, and love) than males, who use more action verbs (e.g., say, speak, and tell) than females. Although this study was performed on texts in Spanish, the authors believe that their methodology can be applied to other languages, for which resources such as emotion dictionaries and tools such as PoS taggers are available.

Fatima et al. (2017) developed a multilingual (English and Roman Urdu) approach to age and gender detection of Facebook authors by (1) constructing a multilingual dataset, (2) manually constructing a bilingual dictionary for translating Roman Urdu into English, (3) creating a model of current state-of-the-art AP techniques by using both content-based features (word and character n-grams) and stylistic-based features, which contain 11 lexical word-based features, 47 lexical character-based features, and 6 vocabulary-richness measures, and (4) comparing and evaluating the behavior of the developed methods on translated and multilingual datasets. The empirical evaluation of their experiment demonstrates that (1) current AF techniques can be used on multilingual datasets (English and Roman Urdu) as well as on monolingual datasets (derived from translation), (2) content-based methods yield greater accuracy than stylistic-based methods for both gender and age detection, and (3) translation of a multilingual dataset to a monolingual dataset does not yield better detection results.

Kocher and Savoy (2017) investigated 24 distance measures (derived from five general function sets) that were used to evaluate the success of AP tasks. In addition, they presented six theoretical properties: zero distance, positivity, symmetry, triangle inequality, frequent feature, and presence of the feature. Thirteen test datasets that were downloaded from the CLEF 2016 evaluation were used to empirically evaluate the effectivity of the 24 distance measures. These test datasets cover four languages (Dutch, English, Italian, and Spanish) and four text genres (blogs, reviews, social media, and tweets) concerning to gender and from 4 to 5 age groups. The experimental results show that (1) the Tanimoto or Matusita distance measures respect all six proposed properties; (2) the Canberra or Clark distance measures present higher effectivity than other measures concerning the AP tasks; and (3) as expected, using a training set that is closely relevant to the test set impacts the overall accuracy achieved.

López-Monroy et al. (2015) proposed a representation method that captures sub-profile-specific and discriminative information. In their method, they present terms using a vector space model that captures discriminative information. The representations of the terms are combined to represent the content of the documents in a low-dimensional non-sparse discriminative space. The authors evaluated the proposed representation's effectiveness on a few social media datasets. They compared their representation to the standard BoW representation on an extensive variety of state-of-the-art AP methods. The results of the experiment show that the proposed representation obtained higher results than most reference methodologies. They also show that the results of the proposed representation are consistent with those of earlier studies on handcrafted attributes for AP.

Gómez-Adorno et al. (2016) introduced a lexical resource for preprocessing social media. The resource is composed of dictionaries of four languages, with 7,259 entries. Each dictionary contains abbreviations, contractions, emoticons, and slang words. They conducted experiments on two datasets composed of Twitter messages downloaded from the PAN 2015–2016 competitions, which consist of 324 and 1,060 tweets, respectively, in four languages. They applied various ML methods (especially LR and SVM) using the Doc2vec neural network (NN) representation method (Le and Mikolov, 2014)). They showed that the use of their resource significantly improves the quality of their representation, for identifying the gender and age of the authors of social media messages. They showed that, in most cases, the results' improvements obtained by using the developed dictionaries are statistically significant.

Soler-Company and Wanner (2017; 2018) showed that deep linguistic features e.g., syntactic features (, PoS-based, and tree-shape) and discourse features (various relation features e.g., Attribution, Background, Cause, Condition, and explanation)) that were extracted by the discourse parser of Surdeanu et al. (2008) achieved state-of-the-art results in gender and author classification while keeping the feature set small. They applied linear kernel SVMs on various literary and blog post datasets. A model using all the features applied on the blog dataset identified the authors' gender with an accuracy of 89.97%, outperforming every presented baseline. Character-based features such as comma, semi-colon, and period usage also obtained a high accuracy result (87.91%). The syntactic features obtained an accuracy of 85.17%, which also outperformed every baseline, showing that there are clear gender-specific patterns. A model containing character-based, discourse-based, dictionary-based, sentence-based, and word-based features applied on the literary dataset identified the authors' gender with an accuracy of 91.78%, significantly outperforming every presented baseline. A combination of discourse and syntactic features obtained an accuracy of 91.46% while syntactic features obtained an accuracy of 90.76%.

Thelwall and Stuart (2019) discovered significant gender differences while exploring millions of comments that were posted to various popular subreddits. Most comments posted to Reddit in general and most its subreddits, in particular, are dominated by males. In the subreddit of news, they found female-oriented topics such as health (doctor, health, medical, and nurse), teachers, and rape. They found that all the sports subreddits are strongly dominated by males. In fitness, which is the sports subreddit least dominated by males they discovered a few female-related terms such as eating (eating, fruit, food, and veggies), females (female, girl, her, and woman), and methods (pilates and yoga). In the subreddit of relationships, where females posted slightly more than half of the comments, females used more terms, e.g., boyfriend, husband, mom, and mother, while males used more terms, e.g., bro, dude, and

man.

Ameer et al. (2019) presented their method for age group and gender detection. Their method was tested on several PAN-AP shared task text datasets (2014 and 2016) written in English. The researchers applied five classical ML methods (LR, J48 (an implementation of C4.5, a decision tree model (Quinlan (2014)), Sequential minimal optimization (SMO) (Platt, 1998), NB, and RF) used on various combinations of feature sets e.g., character n-grams, word n-grams, PoS n-grams, and syntactic PoS n-grams. The best accuracy scores (0.496 for age group and 0.734 for gender) have been obtained by SMO using a combination of word n-grams (n = 1, 2, 3, 4).

In summary, in this section, we have presented an overview of a variety of representative studies conducted over the years 2002–2019 that are related to gender and/or age identification/classification. In the next section, we present summaries of many of these studies. The summaries are presented in two tables (Table 1 and Table 3). Table 1 describes popular non-PAN datasets for age and/or gender profiling tasks along with the type of their documents, classification classes, number of documents, number of words, classification tasks, and links. Table 3 presents the accuracy results obtained by various systems on different popular datasets along with the selected features, the applied ML methods, and the preprocessing methods. After each of these tables, there are summaries for each discussed aspect.

4. Overview of representative datasets

Tables 1 and 2 present various details about many popular non-PAN and PAN datasets for age and/or gender profiling tasks, respectively. Details about other such datasets can be found in Wiegmann et al. (2019).

A summary of Table 1 shows the following findings:

Datasets and type of documents: The first two datasets contain formal written texts such as fiction and non-fiction BNC documents as well as other BNC articles and books from different genres. The next three datasets contain blog posts. The two last datasets contain Russian-language texts. The first one contains texts of different genres. The second one contains previews of fiction books.

Language: English – 5 datasets, and Russian – two datasets.

Classification classes: Gender and/or age –2 datasets; only gender – 4 datasets, only age – 1 dataset.

Number of documents: The first two datasets include a few hundred documents. Three datasets contain thousands of documents and two datasets contain tens of thousands of documents.

Number of words: The average number of words per document varies from several hundred words in the blog datasets to several tens of thousands of words in the BNC datasets.

Classification tasks: Gender only (4 datasets), gender and/or age (2 datasets), and age only (1 dataset).

Links: Unavailable links- 4 datasets. Available links – 3 datasets.

A summary of Table 2 shows the following findings:

Datasets and type of documents: All the datasets contain social media texts. The first two datasets contain blog posts. The other five datasets contain Twitter messages (one of them also contains blog posts, and another one also contains images).

Language: During all the years, there are datasets in English and Spanish; all other years contain datasets in 3 or 4 languages. Among the additional languages are Dutch, Italian, Portuguese, and Arabic.

Classification classes: Gender and/or age – the first 4 datasets; only gender (sometimes with another classification task, which is not age) – the last three datasets.

Number of documents: Most datasets include a few hundred or a few thousand documents for each class of documents. For almost all datasets, the training sub-corpus contain 1.5 or 2 times more documents than the test sub-corpus. In a few datasets, the difference between the sub-corpora is even bigger (up to 10 times more).

Remark: In one dataset, the training sub-corpus contains Twitter messages, while the test sub-corpus contains blog posts.

Table 3 presents the accuracy results obtained by various systems on different popular datasets.

A summary of Table 3 shows the following findings:

Language: The first system worked on a dataset containing documents in English and Spanish. All other systems worked only on English documents.

Classification classes: Gender and/or Age – 4 datasets; only gender – 5 datasets.

Applied ML methods: Exponential Gradient (EG) – 4 first systems, Winnow (versions) – 2 next systems, NB – 2 systems, and SVM – one system. The best applied ML methods according to this table are EG and NB. However, it is important to point out that the presented results are relatively old (years 2002–2010) and many modern ML and DL methods were not tried or even did not exist then.

Preprocessing: Five systems did not apply any preprocessing method. One system applied stopword removal and found that the results without removing the stopwords were more accurate. One system did not report the application of any preprocessing method. Two systems ignored formatting issues and removed non–English text.

Features: Five systems used function words, seven systems used various PoS features, and four systems used content features (usually most frequent word unigrams).

Accuracy results: The results depend on the tested datasets, applied ML methods, and features. The gender accuracy results vary from 61.69% to 89%. The age accuracy results vary from 76.2% to 80%.

5. Overview of PAN competitions and winners

In this section, we present several tables that summarize all previous PAN's overview papers about AP tasks between the years 2013 to 2019. For each year, there is a separate detailed overview paper. Therefore, we will not summarize these PAN tasks again.

The approaches in each PAN task were analyzed from three viewpoints: features, classification methods, and preprocessing. All PAN's AP tasks dealt with the prediction of authors' gender. However, not all of the PAN's AP tasks dealt with the prediction of authors' age. Some of the tasks dealt with personality traits (PAN15) and language variety identification (PAN17).

In Table 4, we present a general summary for each PAN's tasks with emphasis on the gender and age sub-tasks, which are the subject of this paper.

The main findings presented in Table 4 are: (1) The number of the teams that participated between 2013 and 2018 (except for an exception in 2014) was quite similar (between 21 and 23); (2) At the same time, between 2013 and 2018, there was an increase in the number of notebook papers that were submitted by the participating teams; and (3) In 2019, there was a significant jump in the number of the participating teams and submitted articles.

In Table 5, in contrast to the PAN's overview papers that present summaries for all the participating teams, we present a summary only for the winners of each PAN task.

Rosso et al. (2019) discussed the evolution of the PAN lab on digital text forensics from 2009 to 2017. PAN focuses on the assessment of various tasks from digital text forensics to develop large-scale, standardized benchmarks, and to assess the state of the art. The authors introduce the evolution of three shared tasks: author identification, AP, and plagiarism detection. The authors carefully noted that maximizing the results on the PAN datasets is not the main goal for the researchers, however, it is very important that PAN will promote reproducibility by requiring program submissions and encouraging competitors s to submit their open-source code.

A summary of Table 5 shows the following findings:

Applied ML methods: Gender: Linear SVM – 6 systems, LR – 2 systems, RF, a decision tree, and Liblinear – one system each. Age: Linear SVM – 2 systems, RF and a decision tree – one system each. The best

Table 1
Popular non-PAN datasets for age and/or gender profiling tasks.

#	1	2	3	4	5	6	7
Dataset description	Formal Written Texts: a genre-controlled gender corpus taken from the BNC	Formal Written Texts: a genre-controlled gender corpus taken from the BNC	The Blog Authorship Corpus: Age and/or gender blog corpus	Age and/or gender blog corpus	Blog Dataset	The RusPersonality dataset contains different texts from different genres written in Russian.	RusAge: Corpus for Age-Based TC
Documents Type	Fiction and non-fiction BNC documents	BNC documents: articles and books from different genres	Blog entries	Blog entries	Blog posts	Texts written by respondents during various experiments, (e.g. description of pictures and essays on different topics) labelled with information on their authors: age, gender, and results of psychological tests	Previews of fiction books (about 5–10% of the books) written by children and adults
Language(s)	English	English	English	English	English	Russian	Russian
Prominent studies applied to this dataset	Koppel et al. (2002)	Argamon et al. (2003)	Schler et al. (2006), Goswami et al. (2009), Argamon et al. (2009)	Argamon et al. (2007), Mukherjee and Liu (2010)	Mukherjee and Liu (2010) run on their dataset their system as well as other systems	Sboev et al. (2016)	Glazkova et al. (2020)
Classification tasks Classes	Gender: Male (M) /female (F)	For each genre M/F: Equal number of male- and female-authored documents	Age: 10′s: 13–17 (42.7%), 20′s: 23–27 (41.9%) 30′s+: 33–47 (15.5%) Gender: M/F	Age: 10′s (13–17): 8,240 blog posts 20′s (23–27): 8,086 blog posts 30′s+ (33–47): 2,994 blog posts Gender: M/F	Gender: M/F	Gender: M/F	Age: children/ adults
Number of documents	566 documents: 264 fiction documents 302 non-fiction documents	604 documents from BNC	37,478 blogs, 1,405,209 blog entries	19,320 blogs	3100 labeled blogs	556 documents written by 556 respondents. Each document contains description of a picture and a letter to a friend.	Train sub-set: 4,492 previews Test sub-set: 1,000 previews
Number of words/tokens	19,425,120	over 25 million words	295,526,889	over 140 million words	895,960	~300,000	2,709,344
Avg. number of words/tokens per doc.	34,320	above 42,000 words	210.32	7,300 words in each blog	289.02	350	499.33
Distribution number of classes	132 fiction/ female and 132 fiction/male; 151 non-fiction/female and 151 non-fiction/male	604 documents divided into 10 genres. For each genre, there is an equal number of documents.	For each age group, an equal number of male and female blogs	10′s (13–17): 8,240 blog posts 20′s (23–27): 8,086 blog posts 30′s (33–42): 2,994 blog posts	1,588 blog posts (51.2%) were written by men and 1,512 blog posts (48.8%) were written by women.	N/A	Train sub-set: 4,492 previews children – 2,108 adults – 2,384 Test sub-set: 1,000 previews: children – 500 adults – 500
Link	N/A	N/A	http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm	N/A	N/A	http://ruscorpora.ru/old/en/	https://github.com/oldaandozerskaya/age_based_classification

applied ML method according to the PAN winners are different variants of Linear SVM.

Preprocessing: Three systems (most earlier systems) did not apply any preprocessing method. One system did not report any applied preprocessing method. Most recent systems applied various preprocessing methods, e.g., lowercase conversion, replacement of various strings (e.g., URLs, LF characters, and User Mentions).

Features: Most systems used word unigrams and bigrams and character 3–4-5-grams. Several systems used various types of stylistic features, e.g., PoS, quantitative, punctuation characters, and vocabulary richness.

Accuracy results: The results depend on the tested languages, datasets, applied ML methods, and features. The age accuracy results vary from 52% to 82%. The gender accuracy results vary from 52% to 91% (most are less than 85%).

6. Profiling age and gender of authors using deep learning methods

Traditional TC systems often use content-based features (bag-of-words, n-gram) and/or style-based features (e.g., quantitative, orthographic features, PoS tags, and function words). Many of these traditional systems use ML methods such as LR, NB, RF, and SVM. In recent years, the use of deep learning (DL) methods such as multi-layer perceptron (MLP), convolutional neural network (CNN), recurrent neural network (RNN), and LSTM model has gained incredible momentum. This is not the case in the area of age and gender profiling. However, there are several systems in this area that have used DL methods. Figs. 4-

Table 2
PAN datasets for age and/or gender profiling tasks.

#	1	2	3	4	5	6	7
Dataset Name	PAN AP 2013 corpus	PAN AP 2014 corpus	PAN AP 2015 corpus	PAN AP 2016 corpus	PAN AP 2017 corpus	PAN AP 2018 corpus	PAN AP 2019 corpus
Documents Type	blog posts	blog posts and social media, Twitter messages, and hotel reviews	Twitter messages	Training: Twitter messages Test: blog posts	Twitter messages	Twitter messages & images	Twitter messages
Classification Task(s)	Age: 18–24, 25–34, 35–49, 50–64, 65+ Gender: M/F	Age: 18–24, 25–34, 35–49, 50–64, 65+ Gender: M/F	Age: 18–24, 25–34, 35–49, 50+ Gender (and other personality traits)	Age: 18–24, 25–34, 35–49, 50–64, 65+ Gender: M/F	Gender (and (variation of native language)	Gender	Bots and Gender Profiling - Gender of a human in the bot/human identification task
Lang-uage(s)	• English (EN) • Spanish (ES)	• EN • ES Reviews are in EN only	• EN • ES • Dutch (DT) • Italian (IT)	• EN • ES • DT	• EN • ES • Portuguese (PT) • Arabic (AR)	• EN • ES • AR	Bots and Gender profiling • EN • ES
Overview paper	Rangel et al., (2013)	Rangel et al., (2014)	Rangel et al., (2015)	Rangel et al., (2016B)	Rangel et al., (2017)	Rangel et al., (2018)	Daelemans et al. (2019)
Prominent studies applied on this task	Meina et al., 2013, Santosh et al., 2013	López-Monroy et al., 2014Markov et al. (2016)	Alvarez-Carmona et al. (2015); González-Gallardo et al. (2015), Grivas et al. (2015), Markov et al. (2016)	Modaresi et al. (2016) Busger et al. (2016) Markov et al. (2016)	Basile et al. (2017) Martinc et al. (2017)	Daneshvar and Inkpen (2018) Tellez et al. (2018)	Valencia et al. (2019)Pizarro (2019)
Training sub-corpus	EN- 236,600 ES – 75,900 This sub-corpus is balanced by gender and imbalanced by age group	Blog posts: EN- 147 ES-88 Social media: EN-7,746 ES-1,272 Tweets: EN- 306 ES-178 Reviews: M – 2,080 F – 2,080	Age: EN: 152 ES: 110 Gender-F/M EN: 152 ES: 110 DT: 34 IT: 38	Twitter messages Age: 18–24, 25–34,35–49,50–64, 65+ Gender: F/M EN: 426 ES: 250 DT: 384	300 authors (500 tweets per gender and variety, 100 tweets per author)	number of authors EN – 3,000 ES – 3,000 AR – 1,500 For each author, 100 tweets and 10 images.	EN: number of authors (100 tweets per author) Bots: 2,060 Female: 1,030 Male: 1,030 ES: number of authors (100 tweets per author) Bots: 1,500 Female: 750 Male: 750
Test sub-corpus	EN- 25,440 ES – 8,160 This sub-corpus is balanced by gender and imbalanced by age	Blog posts: EN- 78 ES- 56 Social media: EN-3,376 ES-566 Tweets: EN- 154 ES-90 Reviews: M – 1,084 F – 1,016	Age: EN: 142 ES: 88 Gender-F/M EN: 142 ES: 88 DT: 32 IT: 36	Blogs EN – 78 ES– 56 For each blog, there are up to 25 posts. The blog subcorpus is balanced by gender	Twitter messages 200 authors (500 tweets per gender and variety, 100 tweets per author)	number of authors EN – 1,900 ES – 2,200 AR – 1,000 For each author, 100 tweets and 10 images.	EN: number of authors (100 tweets per author) Bots: 1,320 Female: 660 Male: 660 ES: number of authors (100 tweets per author) Bots: 900 Female: 450 Male: 450
Link to the PAN profiling task	https://pan.webis.de/clef13/pan13-web/author-profiling.html	https://pan.webis.de/clef14/pan14-web/author-profiling.html	https://pan.webis.de/clef15/pan15-web/author-profiling.html	https://pan.webis.de/clef16/pan16-web/author-profiling.html	https://pan.webis.de/clef17/pan17-web/author-profiling.html	https://pan.webis.de/clef18/pan18-web/author-profiling.html	https://pan.webis.de/clef19/pan19-web/author-profiling.html

6 present the number of publications in Google Scholar having "deep"[3] and "author profiling", "age and classification/categorization", and "gender and classification/categorization" in their title in the years 2014–2020[4], respectively.

7. Analysis of Figs. 4-6

After the publication of Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), there were only three papers in 2014–2015 having "gender," "deep," and "classification/categorization" in their title (Fig. 6), with no papers in the other two categories (Figs. 4-5) in these years. In 2017, there was a significant increase in the number of articles. In 2018–2020, in most cases in Figs. 4-6, there is a relative decrease in the number of articles compared to the number of such articles in 2017. At the same time, the numbers in 2018–2020 are higher than those reported in 2016 and below. As mentioned several times, according to our experience, the results of the last two years (2019–2020) are unstable and do not reflect the truth; these results are expected to increase.

In general, it can be said that in 2017 there was a significant revival in the number of articles, Then there was a certain decrease. In any case, the number of articles in recent years is relatively low compared to those engaged in classification using classical ML methods. The issue of deep learning has not yet developed enough in this field classification and probably therefore the results of the DL are relatively low. An increase in

[3] We decided to search for "deep" instead of "deep learning" because "deep" allows retrieval of phrases such as "deep network," "deep neural networks," "deep averaging networks," and "deep multi-task learning.".
[4] Figures 4–6 are updated to 6-Jun-2021.

Table 3
Comparison of accuracy results obtained by different systems on different popular datasets.

System/ article	Koppel et al. (2002)	Argamon et al. (2003)	Mukherjee and Liu (2010)	Krawetz (2006)	Argamon et al. (2007)	Schler et al. (2006)	Goswami et al. (2009)	Yan and Yan (2006)	Mukherjee and Liu (2010)
Dataset	Koppel et al. (2002)	Argamon et al. (2003)	Mukherjee and Liu (2010)	Gender Guesser	1) Argamon et al. (2007) 2) Mukherjee and Liu (2010)	1) Schler et al. (2006) 2) Mukherjee and Liu (2010)	Schler et al. (2006)	1) Yan and Yan (2006) 2) Mukherjee and Liu (2010)	Mukh-erjee and Liu (2010)
Lang-uage	English Spanish	English	English	English	English	English	English	English	English
Classif-ication task(s)	Gender	Gender and/or genre	Gender	Gender	Age and/or gender	Age and/or gender	Age and/or gender	Gender	Gender
ML method	Exponential Gradient (EG)	a version of the EG algorithm (Kivinen 1997), a generalization of the Balanced Winnow algorithm (Littlestone 1988)	Version of the algo-rithm used in (Argamon et. al, 2003)	Version of the algo-rithm used in (Argamon et. al, 2003)	1) BMR WIN	MCRW	NB	NB	SVM Regression with ensemble feature selection (EFS)
Pre-process-ing methods activated on the dataset	None	None	None	Un-known	Formatting was ignored Non–English text was removed	Formatting was ignored No difference between quotes within a blog and the text of the blog itself. Non–English text removed	None	Results with stopwords were higher than results without stopwords	None
Types of features and their numbers	1,081 features: 405 function words, and 676 PoS-Tags n-grams: 500 triples, 100 pairs, and 76 single tags	1,143 features: 467 function words, PoS-tags: 500 most common triples, 100 most common pairs, and all 76 single tags.	Implem-ented features of the algorithm in (Argamon et. al, 2003)	Implem-ented features of the algorithm in (Argamon et. al, 2003)	1000 most frequent words: 323 function words & 677 content words	502 stylistic features (PoS tags, blog words, and hyperlinks) 1000 word unigrams with the highest information gain	**Age:** 35 content words, 52 slang words, and average sentence length, **Gender:** 35 content words with 52 slang words	content words, blog-words results of dictionary-based content analysis, PoS unigrams	23,974 PoS Patterns
Accu-racy (%)	77.3	Gender: Approximately 80	61.69	63.78	Age: 77.4 (BMR) Gender: 80.5 (WIN)	Gender: 80.1 Age: 76.2	Age: 80 Gender: 89	68.75	88.56

Table 4
PAN competitions (about age and/or gender).

PAN's Year	Paper	Task	Languages	Dataset(s)	Number of teams	Number of submitted notebook papers
13	Rangel et al. (2013)	gender & age	English Spanish	blog posts	21	18
14	Rangel et al. (2014)	gender & age	English Spanish	social media, blogs, tweets, and hotel reviews	10	8
15	Rangel et al. (2015)	gender & age (& personality traits)	English, Spanish, Italian, Dutch	tweets	22	21
16	Rangel et al. (2016)	gender & age)Dutch – only gender)	English, Spanish, Dutch	social media, blogs, essays, reviews	22	14
17	Rangel et al. (2017)	gender (& language variety identification)	English, Spanish, Arabic Portuguese	tweets	22	20
18	Rangel et al. (2018)	gender	English, Spanish, Arabic	tweets & images	23	22
19	Daelemans et al. (2019)	Bot or human (gender: male or female)	English Spanish	tweets	56	46

the number of studies in this field may lead to a significant improvement in the results and consequently to an increase in the number of studies/ articles in the field.

The systems that apply deep neural networks (NNs) use another type of feature, i.e., word embedding vectors. The word embedding method maps each word to a real-valued dense vector. The resulting vectors capture semantic and syntactic relations between words. This method enables to measure word relatedness as the distance between two embedding vectors. There are four popular word embedding methods: Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), ELMo

Table 5
Informative details about the systems of PAN winners.

Paper	ML Method (Best method in bold)	Preprocessing	Feature Sub-sets	Task & Language	Best Result (Accuracy)
Meina et al. (2013)	• **Random Forest** • Naive Bayes • Random Tree • linear SVM • SVM with RBF	• Remove most HTML tags • Delete documents, which contain spam words at a rate of 0.1% and above • Word tokenization • Sentence splitting	311 features divided into the following sets: Content-based featuresPoS features – relative frequencies of particular parts of speechtopic-based features estimated using Latent Semantic Analysis (LSA) (Dumais, 2004)	Gender EN Age EN	59.21 64.91
Santosh et al. (2013)	• **Decision tree of classifiers using SVM and MaxEnt**	None	• content-based features – top 40,000 word n-grams • style-based features: n-grams of PoS tags, punctuation symbols, and number of HREF links • topic-based features estimated using LSA	**Gender** ES **Age** ES	 64.73 64.30
Lopez-Monroy et al. (2014)	**Lib-LINEAR classifier**	None	n-Second Order Attributes (SOA) features (document vectors) per profile based on relationships among the most 3,000 frequent terms, documents, profiles, and sub-profiles. SOA (Lopez-Monroy et al., 2013) is a supervised method that builds document vectors in a space of the target profiles	**Gender** 4 corpora: social media, blogs, Twitter, and hotel reviews in both EN & ES	28.95
Alvarez-Carmona et al. (2015)	**Lib-Linear SVM**	None	• Descriptive features, i.e., LSA features & TFIDF features • SOA– most frequent discriminative features (e.g., function words, stopwords, and punctuation marks)	**Gender** EN ES IT DU **Age** EN ES	 78.28 91.00 86.64 91.17 79.60 82.00
Modaresi et al. (2016)	**Logistic regression**	Remove punctuation signs	• Word unigrams and bigrams • Character 4-grams within word boundaries • average spelling error • (only for age task): four punctuation averages: ",","."," !", "?"	**Gender** EN	51.79
Deneva (no paper)	Unknown	Unknown	Unknown	**Gender** ES	73.21
		• Word tokenization using the Natural Language Toolkit (NLTK) (Bird et al. (2009)	• Word/Character n-grams • PoS • Quantitative	**Age** EN **Age** ES	58.97 51.79
Busger et al. (2016)	**Linear SVM**	• Sentence splitting • Convert several tokens to placeholders • Replace all URLs with the string 'URL' • Replace all numbers with 'NUMBER' • HTML mark-up are deleted	• punctuation characters • grammatical correctness • Vocabulary Richness features • Function words • occurrences of the top 2,500 most frequent words for each category		
Basile et al. (2017)	**Linear SVM**	• using the default scikit-learn tokenizer • did not use PoS-tags as features because they dropped the results	combinations of character 3–4-5-grams and word 1–2-grams with tf-idf weighting	**Gender** EN ES AR PT	 82.33 83.21 80.06 84.50
Danes-hvar and Inkpen (2018)	**Linear SVM**	• Replace LF characters • Concatenate all tweets of each author into one string • lowercase conversion • Replace repeated character sequences • Replace URLs • Replace @username mentions • Remove punctuations • Remove stop words	• Word unigrams, bigrams, and trigrams • Character 3–4-5-grams • topic modeling using 300 dimensions of latent semantic analysis (LSA) • Word unigrams and bigrams Character 3–4-5-grams	**Gender** EN **Gender** ES	82.21 82.00
Tellez et al. (2018)	**Linear SVM**	diacritic removal, character duplication removal, punctuation removal, case normalization	• Word unigrams, bigrams, and trigrams • character 1–3-5–7-9-grams • word skip-grams (2, 1), (2, 2), (3, 1)	**Gender** AR	81.70
Valencia et al. (2019)	**Logistic Regress-ion**	Based on Daneshvar and Inkpen (2018): 1. Replace emojis with their textual description 2. Replace URLs with the word "link" 3. Replace User Mentions with the word "usermention" 4. Replace LF characters with the word "linefeed" 5. Replace special characters with text using unicode data library	• Word unigrams, bigrams, and trigrams • Character 3–4-5-grams • Emojis • Special Characters	**Gender** EN	84.32

(continued on next page)

Table 5 (*continued*)

Paper	ML Method (Best method in bold)	Preprocessing	Feature Sub-sets	Task & Language	Best Result (Accuracy)
		6. lowercase conversion 7. Trimmed repeated characters 8. 8All n-grams that occurred in all documents were ignored			
Pizarro (2019)	**Linear SVC**	• The XML files were parsed • 100 tweets of each author were concatenated forming one string, and a custom tag was used to separate the tweets. • Lowercase conversion • Strings are tokenized using NLTK • Each URL, user mention, and hashtag were replaced by one fixed tag respectively based on Daneshvar and Inkpen (2018)	• Word unigrams, bigrams, and trigrams • Character 3-4-5-grams	**Gender** ES	81.72

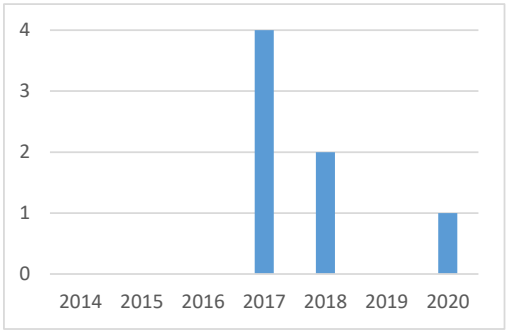

Fig. 4. The number of publications in Google Scholar having "deep" and "author profiling" in their title in the years 2014–2020.

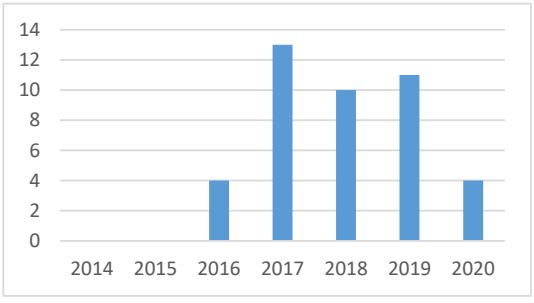

Fig. 5. The number of publications in Google Scholar having "age," "deep," and "classification/categorization" in their title in the years 2014–2020.

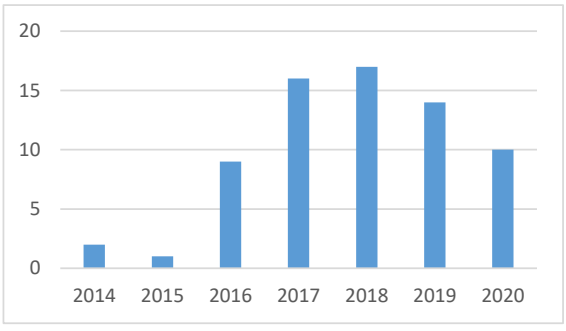

Fig. 6. The number of publications in Google Scholar having "gender," "deep," and "classification/categorization" in their title, in the years 2014–2020.

(Peters et al., 2018), and BERT. The first two methods are context-independent (i.e., they do not take into account the word position in a sentence). They produce only one embedding vector for each word, combining all the different senses of the word into one vector. The last two methods are context-dependent (i.e., they do take into account the word position in a sentence). They produce several embedding vectors for a word that appears several times in the same sentence.

Table 6 presents general information about seven relevant systems, including the datasets they worked on, the DL methods they used, the features that were applied, and the results that they obtained.

A summary of Table 6 shows the following findings:

Language(s): Two studies on datasets in 4 languages: English, Spanish, Arabic, and Portuguese. Two systems worked on Russian documents, one system on Turkish documents, and three systems on two datasets of Twitter messages in Arabic.

Classification classes: Gender: 4 systems. Gender and age: three systems.

Applied ML methods: CNN: 3 systems, RNN: one system, BERT: one system. The best applied ML method according to this table is CNN. However, this method was also the most applied on. Other DL methods were tried very little or not tried at all.

Accuracy results: The accuracy results of the gender classification task vary from 65% to 85%, depending on the tested languages, datasets, ML methods, and features. The rates of the age classification task on the two Arabic datasets vary from 22% to 55%.

An overview of the systems presented in Table 6 is as follows. Sboev et al. (2016) performed comparative research of ML methods for two tasks: identification of the author's gender, and the text's sentiment. They used RusPersonality - a dataset containing texts written in Russian labeled with details on their authors (e.g., age and gender) for the gender-identification task, with the data of the SentiRuEval competition that are dedicated for the sentiment analysis task. For gender identification, they used 141 features of four different groups: 13 PoS tag features, 60 syntactic features, 37 emotion-based features (e.g., anxiety, discontent), and 31 part-of-speech sequences and special characters (e.g., the number of emoticons). They compared nine different ML methods. Two of them (CNN + MLP and CNN + LSTM) are complicated topologies of Artificial Neural Network (ANN) and include 11 layers using various word2vec models. The 9th and 11th layers contain ten and two hidden neurons respectively. This study showed that the two complicated NN models obtained the best results with an accuracy result that is around the state-of-the-art (0.86+/-0.03 for CNN + LSTM and 0.83+/-0.05 for CNN + MLP, both using grammatical information as feature selection technique), but the drawback of such models is the difficulty in assessing the features. For the other models, Rectified Linear Unit (ReLU, 1 Hidden Layer) was the most efficient algorithm, with an accuracy of 0.74+/-0.05 using the imp_quarter feature selection

Table 6
General information about several systems that used DL methods for age and/or gender profiling.

Paper/ System	Dataset/Language(s)	DL Methods	Results
Sboev et al. (2016)	The RusPersonality dataset - texts are written in labeled with details about their authors	CNN + MLP and CNN + LSTM that include 11 layers using various word2vec models	**Gender: Accuracy** 0.86+/-0.03 for CNN + LSTM 0.83+/-0.05 for CNN + MLP
Schaetti (2017)	PAN CLEF 2017 datasets of tweets, one per different language (English, Spanish, Portuguese, and Arabic)	CNN for letter bi-grams of 6 layers: 10·5·5 kernel, 20·5·5 kernel, drop-out, 2 linear layers (ReLU), and softmax.	**Gender: Accuracy** English: 78% Spanish: 72.3% Arabic: 75% Portuguese: 85%
Kodiyan et al. (2017)	PAN CLEF 2017 datasets of tweets, one per different language (English, Spanish, Portuguese, and Arabic)	A bi-GRU + Attention model of 5 layers: Word2vec embeddings, 2 bidirectional RNN layers with a Gated Recurrent Unit (GRU), Attention Mechanism, and softmax.	**Gender: Accuracy** English: 78.88% Spanish: 72.17% Arabic: 71.50% Portuguese: 78.13%
Sboev et al. (2018)	Four datasets are written in Russian • The RusPersonality dataset. • GI cs • RusPersonality (RusPer) • RusProfiling	Various versions of CNN with 128 neurons and activation function = Relu. Various versions of Stacked Bidirectional LSTM with network topology of CNN: 30 neurons, activation function = Relu, and window size = 2 or 3.	**Gender: F1-Score** The Gradient Boosting ML method using 3-to-8-character n-grams: 0.64better than the results obtained by different versions of CNN and LSTM (the results of the DL methods were not explicitly presented).
Yildiz (2019)	News in Turkish written by 138 males and 138 females. Ten articles for each one. Total of 2,760 articles	CNN model with word embeddings obtained by Word2vec and RNN	**Gender: F1-Score** A few traditional ML methods have proven to be significantly more efficient and successful than various versions of CNN and RNN. stochastic gradient descent (SGD) 0.91, MLP/Doc2vec 0.84, k-NN/Word2vec 0.80, k-NN/Glove 0.74, CNN/ Word2vec 0.69, RNN 0.58
Zhang and Abdul-Mageed (2019)	The dataset of the APDA shared task (Rangel et al., 2019) – Tweets in Arabic (100 tweets for each user) Training set – 2,250 users, 225,000 tweets. Test set – 720 users, 720,000 tweets.	An existing multi-lingual BERT-based model	**Gender: Accuracy** 81.67% **Age: Accuracy** 54.72%
Suman et al. (2019)		An LSTM model with three layers with different activation keys using a word embedding matrix based on Word2Vec and hand-crafted feature vectors	**Gender: Accuracy** 66.25% **Age: Accuracy** 22.22%
Abdul-Mageed et al. (2019)	2.4 M Tweeter messages in Arabic covering 11 Arabic regions. For each region, there are 100 Twitter users with at least 2000 Tweets. Each user was annotated with gender and age within three categories (under 25, 25 to 34, above 35).	The authors used an existing multi-lingual BERT-based model (12 layers, 768 hidden units, 12 attention heads, and 110 M parameters). They applied it with a maximum number of 30 words in a sequence, a batch size of 32, a learning rate of $2{\cdot}10^{-5}$, and 15 epochs.	**Gender: Accuracy** 65.30% **Age: Accuracy** 51.42%
López-Santillán et al. (2020)	The gender and age datasets of Pan during the years 2013–2018	The authors used genetic programming that generates a document embeddings method, which is a weighted average of various word embedding methods, e.g., BERT, fastText, and word2vec.	<table><tr><th>Dataset</th><th colspan="2">Gender</th><th colspan="2">Age</th></tr><tr><th></th><th>Acc</th><th>F1</th><th>Acc</th><th>F1</th></tr><tr><td>PAN-13</td><td>0.6209</td><td>0.6181</td><td>0.6662</td><td>0.4515</td></tr><tr><td>PAN-14 social</td><td>0.6084</td><td>0.6076</td><td>0.4358</td><td>0.3291</td></tr><tr><td>PAN-14 blogs</td><td>0.7214</td><td>0.7146</td><td>0.5549</td><td>0.2630</td></tr><tr><td>PAN-14 Twitter</td><td>0.7804</td><td>0.7801</td><td>0.5556</td><td>0.3212</td></tr><tr><td>PAN-14 Reviews</td><td>0.6827</td><td>0.6826</td><td>0.3167</td><td>0.1828</td></tr><tr><td>PAN-15</td><td>0.9100</td><td>0.9099</td><td>0.8343</td><td>0.7613</td></tr><tr><td>PAN-16</td><td>0.6882</td><td>0.6853</td><td>0.5614</td><td>0.3235</td></tr><tr><td>PAN-17</td><td>0.8087</td><td>0.8086</td><td>--</td><td>-</td></tr><tr><td>PAN-18</td><td>0.8047</td><td>0.8046</td><td>-</td><td>-</td></tr></table>
Kowsari et al. (2020)	An unknown number of Tweeter messages in English. The training set contains messages written by 1,800 females and 1,800 males and the test set contains messages written by 1,200 females and 1,200 males.	The authors applied various CNN models and various random multi-model DL (RMDL) models using TF-IDF and Glove and then a majority vote scheme was used to make the final decision.	**Age: Accuracy** 86.33%
Das and Paik (2021)	Four gender datasets are written in English: 1) CoNLL-g dataset derived from the CoNLL 2003 NER Shared Task data, 2) Wiki-g dataset created from the Wikigold corpus, 3) IEER-g dataset taken from the NEWSWIRE development test data from the NIST 1999 IE-ER Evaluation corpus, 4) Textbook-g dataset – 71 textbooks: 15 textbooks from a NER dataset, and 56 textbooks downloaded from three countries.	a bidirectional cascading transformer (BiCTransformer, BiCTr) method, which contains two couples of transformers (forward and backward in each couple). The first couple is trained on the NER task, and the second is trained on the gender task. Each transformer contains eight encoder blocks and eight decoder blocks. In addition, they applied a combiner layer that contains 2 linear sub-layers, 1 position invertor, 1 position-wise addition, and 1 softmax layer.	**Gender:** BiCTr outperforms two commercial APIs and five supervised DL methods using 5-fold cross-validation: For all four datasets, BiCTr obtained the highest F1-Score (CoNLL-g: 0.89 ± 0.02; Wiki-g: 0.89 ± 0.03; IEER-g: 0.83 ± 0.04; and Textbook-g: 0.86 ± 0.03). BiCTr obtained the highest accuracy result for three out of the four datasets (CoNLL-g: 0.88 ± 0.02; IEER-g: 0.80 ± 0.02; and Textbook-g: 0.82 ± 0.05). In the Wiki-g dataset, the highest accuracy result (0.85 ± 0.08) was obtained by a commercial name-based gender detection API. BiCTr obtained the second highest accuracy result (0.84 ± 0.02).

technique.

An interesting and challenging task is gender detection for texts with gender deception. Sboev et al. (2018) dealt with this task for texts written in Russian. The authors applied several classical ML methods (e.g., SVM, Gradient Boosting, and Decision Tree) as well as deep ML methods (e.g., CNN and LSTM) on four datasets written in Russian: the RusPersonality dataset, GI cs, RusPersonality (RusPer), and RusProfiling. The best result, an F1-score of 0.64 has been obtained by the Gradient Boosting ML method using 3-to-8-character n-grams according to their TF-IDF values. This result was higher than the results obtained by other ML methods including the CNN and LSTM ML methods (the results of the DL methods were not explicitly presented).

Schaetti (2017) proposed a method for AP using TF-IDF, and a DL model to predict the gender and language variety of Twitter authors. He applied his method to the AP task of the PAN CLEF 2017 campaign. The dataset used in the campaign consisted of four collections of tweets, one per different language (English, Spanish, Portuguese, and Arabic). For the DL model, he applied a matrix of letter bi-grams with punctuation marks, beginning bi-grams, and ending bi-grams. Such a matrix, which represents an author, is the input for a CNN, which consists of 6 layers. Applying this strategy, he determined that his DL model obtained the highest accuracy on gender classification. The accuracy results of his CNN-based algorithm for gender classification were significantly better than those using the TF-IDF values for all four languages, as follows: Spanish (72.3% vs. 64.9%), Arabic (75% vs. 68.8%), English (78% vs. 68%), and Portuguese (85% vs. 73.1%).

Kodiyan et al. (2017) presented their DL method for the AP Shared Task at PAN 2017 (classifying language variety and gender of a Twitter user based only on their Twitter messages). The dataset contains tweets in four different languages. They preprocessed the tweets by converting them with TweetTokenizer to tokens and then mapping them with a token-ID, where each ID points to a vector representation of a specific token. The tokens are represented by pre-trained word embeddings. For Spanish and English, they used embeddings generated by word2vec. For Portuguese and Arabic, they used pre-trained embeddings that were trained on Wikipedia (Bojanowski et al., 2017). Their model contains a bi-directional RNN implemented with a GRU combined with an Attention Mechanism and had 4 layers in total: Embedding, GRU, Attention, and Softmax. The bi-GRU + Attention model obtained the best results on both classification tasks (language variety and gender). The CNN model outperformed the bi-RNN model in Spanish and Portuguese, and the bi-RNN outperformed the CNN model in English and Arabic. overall, the bi-RNN with an accuracy of 75.67% exceeds the CNN by 1.45% on average over the four languages.

The accuracy results of Schaetti (2017) (EN: 78, ES: 72.3, AR: 75, and PT: 85) and Kodiyan et al. (2017) (EN: 78.88, ES: 72.17, AR: 71.50, and PT: 78.13), described above and in Table 6, that took part at the PAN at CLEF 2013 were significantly lower (except on case) than the results of the best system at PAN at CLEF 2013 (Basile et al., 2017) described in Table 6, which applied a linear SVM with combinations of character and word n-grams (EN: 82.33, ES: 83.21, AR: 80.06, and PT: 84.50).

An interesting and challenging task is gender detection for texts with gender deception. Sboev et al. (2018) dealt with this task for texts written in Russian. The authors applied several classical ML methods (e.g., SVM, Gradient Boosting, and Decision Tree) as well as deep ML methods (e.g., CNN and LSTM) on four datasets written in Russian: the RusPersonality dataset, GI cs, RusPersonality (RusPer), and RusProfiling. The best result, an F1-score of 0.64 has been obtained by the Gradient Boosting ML method using 3-to-8-character n-grams according to their TF-IDF values. This result was higher than the results obtained by the other ML methods including the CNN and LSTM ML methods (the results of the DL methods were not explicitly presented).

Yildiz (2019) manually collected a dataset from several Turkish news websites. The dataset contains 2,760 articles written in Turkish that were composed by 138 males and 138 females (10 articles for each one). Yildiz applied various ML methods to classify the gender of the authors. A few traditional ML methods have proven to be significantly more efficient and successful than various versions of CNN and RNN. The F1-Scores were as follows: stochastic gradient descent (SGD) 0.91, MLP/Doc2vec 0.84, k-NN/Word2vec 0.80, k-NN/Glove 0.74, CNN/Word2vec 0.69, and RNN 0.58.

Zhang and Abdul-Mageed (2019) present their models for detecting gender, language variety, and age, as part of the Arabic AP and deception detection shared task (APDA) (Rangel et al., 2019). The dataset of the APDA contains Tweets in Arabic (100 tweets for each user) that were divided into a training set – 225,000 tweets (posted by 2,250 users) and a test set – 720,000 tweets (posted by 720 users). Their submitted accuracy results on the gender and age test sets (81.67%% and 54.72, respectively) were obtained by using a multi-lingual BERT-based model[5] (12 layers, 768 hidden units, 12 attention heads, and 110,000,000 parameters) (Devlin et al., 2018). They fine-tuned the model with a maximum number of 30 words in a sequence, a batch size of 32, a learning rate of $2{\cdot}10^{-5}$, and 15 epochs.

Another team that applied a DL method in the APDA task is Suman et al. (2019) who used an LSTM model with three layers with different activation keys using a word-embedding matrix, based on Word2Vec and vectors that contained hand-crafted features (e.g., emoji and word counts, and special characters). In their better model (model-II), they obtained the following accuracy results on the gender (66.25%) and age (22.22%) on the test sets.

Abdul-Mageed et al. (2019) worked on the Arab-Tweet (Zaghouani and Charfi, 2018) dataset. This dataset contains 2.4 M Tweeter messages in Arabic covering 11 Arabic regions from 17 different countries, where each region contains 100 users with at least 2,000 tweets for each user. Human annotators annotate each user with gender and age within three categories (under 25 years old, between 25 and 34, and above 35). The best accuracy results on the gender and age test sets (65.30% and 51.42, respectively) were obtained by the multi-lingual BERT-based model (Devlin et al., 2018) mentioned above. They fine-tuned the model with a maximum sequence size of 30 words, a batch size of 32, a learning rate of $2e^{-5}$, and 15 epochs.

Rangel et al. (2019) report on the Arabic AP and deception detection shared task. In this task, 13 teams have participated, submitting 28 runs. Most systems applied classical ML methods (e.g., SVM, Multinomial Naive Bayes (MNB), LR, and RF) and only two teams applied DL methods: Zhang and Abdul-Mageed (2019), who used BERT pre-trained on Wikipedia, and Suman et al. (2019) who used an LSTM model.

In the APDA task, the best accuracy result for gender classification (81.94%) has been obtained by the MagdalenaYVino team[6] who applied MNB using a combination of word uni-grams and emoticon 2–3-grams. The best accuracy result for age identification (62.50%) has been achieved by Sun et al. (2019), who applied LR trained with a combination of character 2–3-4–5-grams and word unigrams. These results (especially the age result), which have been obtained by classical ML methods (MNB and LR), are much better than the results of the two systems that used DL methods (Zhang and Abdul-Mageed, 2019; Suman et al., 2019).

López-Santillán et al. (2020) applied a genetic programming technique that generates a new document embeddings method, which is a weighted average of various word embedding methods, e.g., BERT, fastText, and word2vec. They evaluated their method on nine datasets by predicting various personal authors' features, e.g., age, gender, and personality traits. A comparison of the results of their method versus the results of state-of-the-art methods shows that their method was at the supreme quarter in all tested datasets. The 28 obtained results are presented in Table 6. The authors also introduced a new feature called Relevance Topic Value, which computes the importance of each term based on the themes and words used by people. The use of this feature helps to improve the classification ability in various AP tasks.

[5] https://github.com/google-research/bert/blob/master/multilingual.md.

[6] No available paper.

Kowsari et al. (2020) applied gender classification on a gender-annotated dataset of Twitter messages written in English. This dataset was generated from the datasets presented in Rangel et al. (2013) and Rangel et al. (2015). The training set contains messages written by 1,800 females and 1,800 males and the test set contains messages written by 1,200 females and 1,200 males[7]. The authors applied various CNN models and various random multi-model DL (RMDL) models using TF-IDF and Glove and then a majority vote scheme was used to make the final decision. The obtained results were accuracy of 0.8633 and an F1-Score of 0.8583.

Das and Paik (2021) suggested a bidirectional cascading transformer (BiCTransformer, BiCTr) method to detect the gender of named entities in text. Their method is composed of two couples of transformers. The first transformers (forward and backward) are trained on the NER task, and the second transformers (forward and backward) are trained on gender tagging. Each transformer contains eight encoder blocks and eight decoder blocks. In addition, they applied a combiner layer that contains 2 linear sub-layers, 1 position invertor, 1 position-wise addition, and 1 softmax layer. The authors generated four gender datasets written in English derived from Named-entity recognition (NER) datasets: 1) CoNLL-g dataset derived from the CoNLL 2003 NER Shared Task data, 2) Wiki-g dataset created from the Wikigold corpus, 3) IEER-g dataset taken from the NEWSWIRE development test data from the NIST 1999 IE-ER Evaluation corpus, 4) Textbook-g dataset − 71 textbooks: 15 textbooks from a NER dataset, and 56 textbooks downloaded from three countries. Regarding gender detection, their method outperforms two commercial APIs and five supervised DL methods using 5-fold cross-validation as follows. For all four datasets, BiCTr obtained the highest F1-Score (CoNLL-g: 0.89 ± 0.02; Wiki-g: 0.89 ± 0.03; IEER-g: 0.83 ± 0.04; and Textbook-g: 0.86 ± 0.03). BiCTr obtained the highest accuracy result for three out of the four datasets (CoNLL-g: 0.88 ± 0.02; IEER-g: 0.80 ± 0.02; and Textbook-g: 0.82 ± 0.05). In the Wiki-g dataset, the highest accuracy result (0.85 ± 0.08) was obtained by the GRize[8] method, which is a commercial name-based gender detection API. BiCTr obtained the second-highest accuracy result (0.84 ± 0.02).

8. Summary, Conclusions, and future research

In this paper, we present an historical overview of the field with several significant leaps. AP in general and especially age and gender classification has become popular in NLP competitions and applications. Fig. 1 shows a dramatic increase in 2013 in the number of publications having "author profiling" in their title. This is probably due to the establishment of the first AP task at PAN in 2013. Figs. 2-3 show that there is a fairly steady gradual increase in the number of articles dealing with two other categories "age and classification/categorization" and "gender and classification/categorization" in their title in most years. Figs. 4-6 show another relatively high and sharp increase in 2016 and especially 2017 in the number of articles dealing with deep learning and ("author profiling" or (age and "classification/categorization") or (gender and "classification/categorization")).

The analysis of the various age and gender datasets shows that most age and gender datasets contain blog posts or Twitter messages written in English, Spanish or Arabic. There are also several datasets written in Dutch, Italian, Portuguese, Turkish, and Russian. The main problems with these datasets are that there is inconsistency and no uniformity in the datasets concerning to the types and numbers of documents they contain, and their division into training sets, dev sets, and test sets. Furthermore, there is also inconsistency in the types of the applied preprocessing methods, and the quality measures that are used to evaluate the classification results.

An interesting finding is that the best age accuracy results (for many tested languages, datasets, applied ML methods, and features) vary from 52% to 82%. The gender accuracy results vary from 52% to 91% (most are less than 85%). These results are quite surprising. We would expect to see higher results when it comes to seemingly relatively simple types of classification especially by gender (only 2 categories) when a large number of teams have competed over the years.

Another interesting finding that repeats itself in various classification tasks such as the PAN at CLEF 2013 task (Rangel et al., 2013), the comparison between ML methods presented by Yildiz (2019), and the APDA task described by Rangel et al. (2019), is that classical ML methods (e.g.; SVM, LR, and RF) are currently better than DL methods for age and gender classification tasks.

Possible explanations for this finding are: (1) the classical methods have been explored and tried much more than the DL methods over the years and on the various datasets; (2) DL methods require large amounts of data to train (Gupta and Gupta, 2019; Brauwers and Frasincar, 2021), which is not the case in most above-mentioned datasets; (3) DL methods are computationally intensive and therefore need advanced computational resources (Montesinos-López et al., 2018; Ayoub et al., 2021); and (4) Optimization of DL methods requires extensive experiments using different combinations of number of layers, number of units, and number of epochs as well as other parameters (Montesinos-López et al., 2018).

Most classical systems used word unigrams and bigrams and character 3–4-5 g. Several systems also used several types of stylistic features. While many earlier systems did not apply preprocessing methods, most recent systems applied several preprocessing methods, e.g., lowercase conversion and replacement of various strings (e.g., URLs, LF characters, and User Mentions).

The answer to the question what are the most successful feature combinations is not unequivocal and varies from study to study depending on various variables, e.g., datasets, ML methods, and feature combinations that have been tested. In almost all the reviewed studies, the best feature combinations (of those tried) were identified in all the reviewed studies. However, in most cases, the authors did not explain why really these feature combinations were the most successful for the discussed classification tasks. In the following paragraphs, we will try to indicate several insights that have emerged in some of the studies.

Content words were found as successful in many age and/or gender classification studies. The reason for that is simply because both different age categories and different gender categories have different areas of interest and therefore they use words from different content categories with different frequencies. For instance: Argamon et al. (2007) found that males use more frequently content words that are associated with Business, Internet, Politics, and Religion, while females use more frequently content words that are associated with At Home, Conversation, Fun, Romance, and Swearing. Argamon et al. (2007) found also that The use of content words that are associated with Business, Family, Internet, Politics, and Religion increases with the increase in age, while the use of words associated with AtHome, Conversation, Fun, Music, Romance, School, and Swearing significantly decreases with the increase in age. Rangel and Rosso (2016) found that "younger people tend to write more about many different disciplines such as physics, linguistics, literature, metrology, law, medicine, chemistry and so on, maybe due to the fact that this is the stage of life when people mostly speak about their homework.".

Also, some of the stylistic features were found as successful in many age and/or gender classification studies. several examples are as follows: Pennebaker et al. (2001) observed that the number of determiners and prepositions usage was increased with age, while the number of negations and pronouns usage were decreased with age. Argamon et al. (2007) found that males use more frequently prepositions and articles, while females use more frequently auxiliary verbs, conjunctions, and personal pronouns. The authors also found that Stylistic features measured by frequencies of grammatical (PoS) tags show that the usage

[7] The authors did not provide the number of Twitter messages included in the dataset.

[8] https://genderize.io/.

of auxiliary verbs, conjunctions, and personal pronouns significantly decreases with age, while the use of prepositions and articles significantly increases with the increase in age. Newman et al. (2008) observed that females use more adjectives and adverbs than male authors. Females are more likely to include verbs, negations, pronouns, words related to home, friends, family, and various emotional words. Males tend to use more articles, prepositions, numbers, and longer words.

Combinations of content features (e.g., word unigrams with the highest information gain) and stylistic features (e.g., function words, hyperlinks, non-dictionary words, and parts-of-speech) were found as the most useful in various age and/or gender classification studies (e.g., Schler et al., 2006; Argamon et al., 2007).

There are various ideas for future research that are relevant to the nature of AP, especially in social texts, e.g., (1) Applying various combinations of preprocessing methods to "clean" the texts from "noise" and improve their quality for TC tasks (HaCohen-Kerner et al., 2020); (2) Using skip character n-grams as general features to overcome problems, e.g., noise and sparse data (HaCohen-Kerner et al., 2017), and (3) applying acronym disambiguation (HaCohen-Kerner et al., 2008B; HaCohen-Kerner et al., 2010C).

General future research proposals include (1) constructing datasets written in various languages that are consistent concerning to the number of documents they contain and their types, and the division into training sets, dev sets, and test sets; (2) implementing TC experiments using both various types of feature types (content-based, stylistic, and word embedding), preprocessing methods, feature filtering, parameter tuning, and measures; (3) adaptation of models that are successful on a specific dataset/domain to another (other) dataset/domain(s); (4) classification of texts containing virtual age and/or virtual gender provided by the authors; and (5) improving DL models based on extensive experiments using various parameters as mentioned above.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

My deepest thanks to Prof. Walter Daelemans for his wise advices during various stages of this paper. I am also grateful to Netanel Sadeh, my student in the past, who helped me with several papers that are related to author profiling using deep learning methods. The author acknowledges partial financial support from the Jerusalem College of Technology (Lev Academic Center) and the COST Action CA16204 "Distant Reading for European Literary History."

References

Abdul-Mageed, M., Zhang, C., Rajendran, A., Elmadany, A., Przystupa, M., & Ungar, L. (2019). Sentence-Level BERT and Multi-Task Learning of Age and Gender in Social Media. arXiv preprint arXiv:1911.00637.

Alvarez-Carmona, M. A., López-Monroy, A. P., Montes-y-Gómez, M., Villasenor-Pineda, L., & Jair-Escalante, H. (2015). INAOE's participation at PAN'15: Author profiling task. Working Notes Papers of the CLEF, 103.

Ameer, I., Sidorov, G., & Nawab, R. M. A. (2019). Author profiling for age and gender using combinations of features of various types. *Journal of Intelligent & Fuzzy Systems, 36*(5), 4833–4843.

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text & Talk, 23*(3), 321–346.

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2007). Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday, 12*(9).

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM, 52*(2), 119–123.

Ayoub, J., Yang, X. J., & Zhou, F. (2021). Combat COVID-19 infodemic using explainable natural language processing models. *Information Processing & Management, 58*(4), Article 102569.

Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., & Nissim, M. (2017). N-GrAM: New Groningen Author-profiling Model—Notebook for PAN at CLEF 2017. In CEUR Workshop Proceedings (Vol. 1866).Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146.

Brauwers, G., & Frasincar, F. (2021). A Survey on Aspect-Based Sentiment Classification. ACM Computing Surveys (CSUR).

Brennan, M., Afroz, S., & Greenstadt, R. (2012). Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC), 15*(3), 1–22.

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). In *Discriminating gender on Twitter* (pp. 1301–1309). Association for Computational Linguistics.

Busger, op Vollenbroek, M., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., Haagsma, H., & Nissim, M. (2016). Gronup: Groningen user profiling. In Working Notes of CLEF, CEUR Workshop Proceedings (pp. 846-857).

Cheng, N., Chen, X., Chandramouli, R., & Subbalakshmi, K. P. (2009). Gender identification from E-mails. CIDM, 9, 154-158.

Cheng, N., Chandramouli, R., & Subbalakshmi, K. P. (2011). Author gender identification from text. *Digital Investigation, 8*(1), 78–88.

Chopra, S., Sawhney, R., Mathur, P., & Shah, R. R. (2020, April). Hindi-English Hate Speech Detection: Author Profiling, Debiasing, and Practical Perspectives. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 01, pp. 386-393).

Coupland, N. (2007). *Style: Language variation and identity*. Cambridge University Press.

Daelemans, W., Kestemont, M., Manjavacas, E., Potthast, M., Rangel, F., Rosso, P., Potthast M., Rangel, F., Paolo Rosso4, Specht G., Stamatatos E., Stein Benno, Tschuggnall M., Wiegmann, M., & Zangerle E. ((2019, September). Overview of PAN 2019: bots and gender profiling, celebrity profiling, cross-domain authorship attribution and style change detection. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 402-416). Springer, Cham..

Daneshvar, S., & Inkpen, D. (2018, September). Gender identification in twitter using n-grams and lsa. In Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric Sanjuan, Linda Cappellato, and Nicola Ferro, editors, Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), September 2018.

Das, S., & Paik, J. H. (2021). Context-sensitive gender inference of named entities in text. *Information Processing & Management, 58*(1), Article 102423.

de Vel, O. Y., Corney, M. W., Anderson, A. M., & Mohay, G. M. (2002). Language and gender author cohort analysis of e-mail for computer forensics.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805.

Dumais, S. T. (2004). Latent semantic analysis. *Annual review of information science and technology, 38*(1), 188–230.

Emmery, C., Kádár, Á., & Chrupała, G. (2021). Adversarial Stylometry in the Wild: Transferable Lexical Substitution Attacks on Author Profiling. arXiv preprint arXiv: 2101.11310.

Eke, C. I., Norman, A. A., Shuib, L., & Nweke, H. F. (2019). A survey of user profiling: State-of-the-art, challenges, and solutions. *IEEE Access, 7*, 144907–144924.

Foong, E., & Gerber, E. (2021). May). Understanding Gender Differences in Pricing Strategies in Online Labor Marketplaces. In *In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16).

Genkin, A., Lewis, D. D., & Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics, 49*(3), 291–304.

Glazkova, A., Egorov, Y., & Glazkov, M. (2020). A Comparative Study of Feature Types for Age-Based Text Classification. arXiv preprint arXiv:2009.11898.

González-Gallardo, C. E., Montes, A., Sierra, G., Núnez-Juárez, J. A., Salinas-López, A. J., & Ek, J. (2015). *Tweets Classification using Corpus Dependent Tags*. In CLEF (working notes): Character and POS N-grams.

Gómez-Adorno, H., Markov, I., Sidorov, G., Posadas-Durán, J. P., Sanchez-Perez, M. A., & Chanona-Hernandez, L. (2016). Improving feature representation based on a neural network for author profiling in social media texts. *Computational intelligence and neuroscience, 2016*, 2.

Goswami, S., Sarkar, S., & Rustagi, M. (2009). Stylometric analysis of bloggers' age and gender. *In Third International AAAI Conference on Weblogs and Social Media*.

Grivas, A., Krithara, A., & Giannakopoulos, G. (2015). September). Author Profiling using Stylometric and Structural Feature Groupings. In CLEF (Working Notes). Gupta, S., & Gupta, S. K. (2019). Abstractive summarization: An overview of the state of the art. *Expert Systems with pplications, 121*, 49–65.

Halliday, M. A. K., Matthiessen, C. M., Halliday, M., & Matthiessen, C. (2014). *An introduction to functional grammar*. Routledge.

HaCohen-Kerner, Y., Mughaz, D., Beck, H., & Yehudai, E. (2008A). Words as classifiers of documents according to their historical period and the ethnic origin of their authors. *Cybernetics and Systems: An International Journal, 39*(3), 213–228.

HaCohen-Kerner, Y., Kass, A., & Peretz, A. (2008B). Combined one sense disambiguation of abbreviations. In Proceedings of ACL-08: HLT, Short Papers (pp. 61-64).

HaCohen-Kerner, Y., Beck, H., Yehudai, E., Rosenstein, M., & Mughaz, D. (2010A). Cuisine: Classification using stylistic feature sets and/or name-based feature sets. *Journal of the American Society for Information Science and Technology, 61*(8), 1644–1657.

HaCohen-Kerner, Y., Beck, H., Yehudai, E., & Mughaz, D. (2010B). Stylistic feature sets as classifiers of documents according to their historical period and ethnic origin. *Applied Artificial Intelligence, 24*(9), 847–862.

HaCohen-Kerner, Y., Kass, A., & Peretz, A. (2010C). HAADS: A Hebrew Aramaic abbreviation disambiguation system. *Journal of the American Society for Information Science and Technology, 61*(9), 1923–1932.
HaCohen-Kerner, Y., Ido, Z., & Ya'akobov, R. (2017, September). Stance classification of tweets using skip char ngrams. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 266-278). Springer, Cham.
HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PloS one, 15*(5), Article e0232525. https://doi.org/10.1371/journal.pone.0232525
Holmes, D. I. (1994). Authorship attribution. *Computers and the Humanities, 28*(2), 87–106.
Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. *Literary and linguistic computing, 13*(3), 111–117.
Juola, P., & Baayen, H. (2003). A controlled-corpus experiment in authorship identification by cross-entropy. *In Literary and Linguistic Computing*.
Karami, A., White, C. N., Ford, K., Swan, S., & Spinel, M. Y. (2020). Unwanted advances in higher education: Uncovering sexual harassment experiences in academia with text mining. *Information Processing & Management, 57*(2), Article 102167.
Kibriya, A. M., Frank, E., Pfahringer, B., & Holmes, G. (2004). December). Multinomial naive bayes for text categorization revisited. In *Australasian Joint Conference on Artificial Intelligence* (pp. 488–499). Berlin, Heidelberg: Springer.
Kivinen, J. (1997). Additive versus exponentiated gradient updates for linear prediction. *Information and computation, 132*(1), 1–64.
Koch, T., Romero, P., & Stachl, C. (2020). Predicting age and gender from language, emoji, and emoticon use in WhatsApp instant messages.
Kocher, M., & Savoy, J. (2017). Distance measures in author profiling. *Information Processing & Management, 53*(5), 1103–1119.
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 Evaluation Labs and Workshop–Working Notes Papers, Dublin, Ireland, 11-14 September 2017 (Vol. 1866). RWTH Aachen.
Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. Literary and linguistic computing, 17(4), 401-412.
Koppel, M., Schler, J., & Zigdon, K. (2005, August). Determining an author's native language by mining a text for errors. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (pp. 624–628).
M. Koppel J. Schler S. Argamon J.W. Pennebaker (2006). Effects of age and gender on blogging. Presented at AAAI, Spring Symposium on Computational Approaches to Analysing Weblogs 2006 CA, March Stanford 2006.
Kowsari, K., Heidarysafa, M., Odukoya, T., Potter, P., Barnes, L. E., & Brown, D. E. (2020). Gender detection on social networks using ensemble deep learning. arXiv preprint arXiv:2004.06518.
Krawetz, N. (2006). *Gender Guesser. Hacker Factor Solutions*. Access Date: 12/Jan/2021.
Kucukyilmaz, T., Cambazoglu, B. B., Aykanat, C., & Can, F. (2006). In *Chat mining for gender prediction* (pp. 274–283). Springer.
Kumar, S., Gahalawat, M., Roy, P. P., Dogra, D. P., & Kim, B. G. (2020). Exploring impact of age and gender on sentiment analysis using machine learning. *Electronics, 9*(2), 374.
Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *In International Conference on Machine Learning* (pp. 1188–1196).
Littlestone, N. (1988). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning, 2*(4), 285–318.
López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J., & Pineda, L. V. (2014). Using Intra-Profile Information for Author Profiling. In CLEF (Working Notes) (pp. 1116-1120).
López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., & Stamatatos, E. (2015). Discriminative subprofile-specific representations for author profiling in social media. *Knowledge-Based Systems, 89*, 134–147.
Lpez-Santamara, L., Gomez, J. C., Almanza-Ojeda, D., & Ibarra-Manzano, M. (2019). Age and gender identification in unbalanced social media. In *In 2019 International Conference on Electronics, Communications and Computers (CONIELECOMP)* (pp. 74–80).
López-Santillán, R., Montes-Y-Gómez, M., González-Gurrola, L. C., Ramírez-Alonso, G., & Prieto-Ordaz, O. (2020). Richer Document Embeddings for Author Profiling tasks based on a heuristic search. *Information Processing & Management, 102227*.
Madigan, D., Genkin, A., Lewis, D. D., & Fradkin, D. (2005). Bayesian multinomial logistic regression for author identification. In AIP conference proceedings (Vol. 803, No. 1, pp. 509-516). AIP.
Markov, I., Gómez-Adorno, H., Posadas-Durán, J. P., Sidorov, G., & Gelbukh, A. (2016). In *October). Author profiling with doc2vec neural network-based document embeddings* (pp. 117–131). Cham: Springer.
Meina, M., Brodzinska, K., Celmer, B., Czoków, M., Patera, M., Pezacki, J., & Wilk, M. (2013). Ensemble-based classification for author profiling using various features. Notebook Papers of CLEF.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
Montesinos-López, A., Montesinos-López, O. A., Gianola, D., Crossa, J., & Hernández-Suárez, C. M. (2018). Multi-environment genomic prediction of plant traits using deep learners with dense architecture. G3: Genes, Genomes. *Genetics, 8*(12), 3813–3828.
Mukherjee, A., & Liu, B. (2010). October). Improving gender classification of blog authors. In *In Proceedings of the 2010 conference on Empirical Methods in natural Language Processing* (pp. 207–217).
Nerbonne, J. (2014). The secret life of pronouns. What our words say about us. *Literary and Linguistic Computing, 29*(1), 139–142.
Newmn, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse processes, 45*(3), 211–236.
Noecker, J., Jr, Ryan, M., & Juola, P. (2013). Psychological profiling through textual analysis. *Literary and Linguistic Computing, 28*(3), 382–387.
Pashutan Modaresi, Matthias Liebeck, and Stefan Conrad. Exploring the effects of cross-genre machine learning for author profiling in PAN 2016.
C. Peersman W. Daelemans L. Van Vaerenbergh Predicting age and gender in online social networks 2011 ACM 37 44.
Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and wod count: LIWC 2001. Mahway: Lawrence Erlbaum Associates, 71(2001), 2001.
Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54*, 547–577.
Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us*. New York: BloomsburyPress.
Pennington, J., Socher, R., & Manning, C. D. (2014). October). Glove: Global vectors for word representation. In *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).
Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et al. (2018). Deep contextualized word rpresentations. *In: Proc of NAACL*.
Pizarro, J. (2019). Using N-grams to detect Bots on Twitter. *In CLEF (Working Notes)*.
Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.
Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.
F. Rangel P. Rosso M. Koppel E. Stamatatos G. Inches Overview of the author profiling task at PAN 2013 2013 CELCT 352 365.
Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., et al. (2014). Overview of the 2nd author profiling task at pan 2014. In *In CLEF 2014 Evaluation Labs and Workshop Working Notes Papers* (pp. 1–30).
Rangel, F., Rosso, P., Potthast, M., Stein, B., & Daelemans, W. (2015, September). Overview of the 3rd Author Profiling Task at PAN 2015. In CLEF (p. 2015).
Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., & Stein, B. (2016). Overview of the 4th authorprofiling task at PAN 2016: cross-genre evaluations. In Working Notes Papers of the CLEF 2016 Evaluation Labs.CEUR Workshop Proceedings/Balog, Krisztian [edit.]; et al. (pp. 750-784).
Rangel, F., & Rosso, P. (2016). On the impact of emotions on author profiling. Information processing & management, 52(1), 73-92. 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. Working Notes Papers of the CLEF.
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.
Rangel, F., Rosso, P., Charfi, A., Zaghoani, W., Ghanem, B., & Snchez-Junquera, J. (2019, December). 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. In: CEUR-WS. org, Kolkata, India.
Reddy, T. R., Vardhan, B. V., & Reddy, P. V. (2016). Profile specific document weighted approach using a new term weighting measure for author profiling. *International Journal of Intelligent Engineering and Systems, 9*(4), 136–146.
Rosso, P., Rangel, F., Farías, I. H., Cagnina, L., Zaghouani, W., & Charfi, A. (2018). A survey on author profiling, deception, and irony detection for the arabic language. *Language and Linguistics Compass, 12*(4), Article e12275.
Rosso, P., Potthast, M., Stein, B., Stamatatos, E., Rangel, F., & Daelemans, W. (2019). Evolution of the PAN lab on digital text forensics. In *Information Retrieval Evaluation in a Changing World* (pp. 461–485). Cham: Springer.
Santosh, K., Bansal, R., Shekhar, M., & Varma, V. (2013). Author profiling: Predicting age and gender from blogs. *Notebook for PAN at CLEF*, 119–124.
Sboev, A., Litvinova, T., Voronina, I., Gudovskikh, D., & Rybka, R. (2016). In *Deep Learning Network Models to Categorize Texts According to Author's Gender and to Identify Text Sentiment* (pp. 1101–1106). IEEE.
Schaetti, N. (2017, September). UniNE at CLEF 2017: TF-IDF and Deep-Learning for Author Profiling. In CLEF (Working notes).
Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006, March). Effects of age and gender on blogging. In AAAI spring symposium: Computational approaches to analyzing weblogs (Vol. 6, pp. 199-205).
Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena, 404*, Article 132306.
Soler-Company, J., & Wanner, L. (2017). On the Relevance of Syntactic and Discourse Features for Author Profiling and Identification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers (Vol. 2, pp. 681-687).
Soler-Company, J., & Wanner, L. (2018). On the role of syntactic dependencies and discourse relations for author and gender identification. *Pattern Recognition Letters, 105*, 87–95.
Stefanija, A. P. (2021). Increasing Fairness in Targeted Advertising: The risk of gender stereotyping by job ad algorithms.
Suman, C., Kumar, P., Saha, S., & Bhattacharyya, P. (2019, December). Gender Age and Dialect Recognition using Tweets in a Deep Learning Framework-Notebook for FIRE 2019. In FIRE (Working Notes) (pp. 160-166).
Sun, Y., Ning, H., Chen, K., Kong, L., Yang, Y., Wang, J., & Qi, H. (2019). Author Profiling in Arabic Tweets: An Approach based on Multi-Classification with Word and Character.

Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., & Nivre, J. (2008). In *The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies* (pp. 159–177). Association for Computational Linguistics.

Tellez, E. S., Miranda-Jiménez, S., Moctezuma, D., Graff, M., Salgado, V., & Ortiz-Bejar, J. (2018, September). Gender identification through multi-modal tweet analysis using microtc and bag of visual words. In Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018).

Thelwall, M., & Stuart, E. (2019). She's Reddit: A source of statistically significant gendered interest information? *Information processing & management, 56*(4), 1543–1558.

Valencia, A. I. V., Adorno, H. G., Rhodes, C. S., & Pineda, G. F. (2019). Bots and Gender Identification Based on Stylometry of Tweet Minimal Structure and n-grams Model.

Verhoeven, B., Daelemans, W., & Plank, B. (2016, May). Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 1632-1637).Werlen, L. M. (2015). Statistical Learning Methods for Profiling Analysis. In Proceedings of CLEF.

Wiegmann, M., Stein, B., & Potthast, M. (2019). July). Celebrity profiling. In *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2611–2618).

Yan, X., & Yan, L. (2006). Gender Classification of Weblog Authors. In AAAI spring symposium: computational approaches to analyzing weblogs (pp. 228-230).

Yildiz (2019). A comparative study of author gender identification. Turkish Journal of Electrical Engineering & Computer Sciences, 27(2), 1052–1064.

Zheng, R., Li, J., Chen, H., & Huang, Z. A framework for authorship identification of online messages: Writing-style features and classification techniques Journal of the American society for information science and technology 57 3 2006 378 393.

Zaghouani, W. & Charfi, A. (2018). ArapTweet: A Large Multi-Dialect Twitter Corpus for Gender, Age and Language Variety Identification. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan.

Zhang, C., & Abdul-Mageed, M. (2019). BERT-Based Arabic Social Media Author Profiling. In: Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. CEUR-WS.org, Kolkata, India.