ISSN 2278-3091
Volume 9 No.2, March - April 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse187922020.pdf
https://doi.org/10.30534/ijatcse/2020/187922020

An Empirical combination of Machine Learning models to Enhance author profiling performance

Roobaea Alroobaea
Department of Computer Science, Taif University, Taif 21974, Saudi Arabia
r.robai@tu.edu.sa

ABSTRACT

When a customer connects to the web site of a given corporation, this corporation tries to gather all the data available about him. After which, the customer gets a service suggestion according to his or her personal details such age and gender. Thus, a context-specific choice-making system based on this information is needed to create an effective categorization. Such a choice system will allow corporations to advertise their marketing. This paper aims to tackle the problem of author profiling (AP). By author profiling, we mean the estimation of the two socio-demographic indicators “Age” and “gender” of these users, based solely on their textual productions. To achieve the best possible classification, we adopt a variety of Machine Learning (ML) algorithms like 'Support Vector Machines(SVM)','Random Forest (RF)','Multilayer Perceptrons (MP)','Decision trees (DT)','Naïve Bayes (NB)', k-Nearest Neighbors(KNN) and Deep Learning technique 'Long Short-Term Memory (LSTM)'. The adopted corpus is taken from the PAN-AP-2015 dataset which is in turn obtained from Twitter. In our work, only English language was considered. The results show that each data set (age or gender) have different results when machine learning techniques were applied. This is related to the power of each used technique. Deep learning techniques proved that they will be helpful when the data set are large.

Key words: Profiling, Gender, Age, Machine Learning, Deep Learning.

1. INTRODUCTION

New web technologies have brought to light a variety of interaction media which today has revolutionized production practices and access to content. The user’s post is the most prolific form of expression. It manifests itself as well in discussion threads, comments from various platforms (videos, media, etc.), blog articles, social network posts, tweets, or even opinions and reviews of products or services.

Discussion forums have been one of the most used asynchronous interaction spaces on the web for the past 20 years. Before being joined by blogs and social networks, these platforms allow Internet users to discuss topics of daily life as varied as family, work, education, health, etc. While preserving a certain anonymity, unlike social networks, whose link with other users is more formalized.

Social networks, forums and blogs have become from this point of view a kind of legitimate reference, a spontaneous place of questioning, exchange of experiences, sharing of solutions and tips .Nevertheless, also the place where the we talk about ourselves and our daily and social concerns [1]. In the eyes of advertisers and manufacturers, this information represents a gold mine in view of what it contains in terms of perceptions, expectations, problems or opinions on services, brands or consumer products.

The democratization of expression spaces thanks to new web technologies generates, beyond the need for innovation in the field of content management, the need to exploit this data through innovative models. The mass of unstructured or semi-structured textual data that today represents the flows generated by the social web is of such importance that it is necessary to resort to efficient analysis methodologies.

The interest aroused by the study of new modes of interaction on the web conceals economic, legal and security issues. These issues are of such importance that they bring up new needs in terms of efficiency and robustness in terms of their operation. The diversification of research perspectives has, from this point of view, made it possible to widen the fields of applications, ranging from e-reputation to economic and strategic intelligence, from the classification of documents to the recognition of named entities, via opinion digging or sentiment analysis.

Among these perspectives, the profiling of authors has aroused growing interest in recent years. Like the Recognition of Named Entities [2], this task of text mining is considered today as a field of application in its own right, giving rise to events/contests. In PAN conference proceedings [3],the main areas of exploration relate to socio-demographic indicators such as age, gender, but also domiciliation, marital or

socio-professional status. Other indicators are also explored, namely the level of education, personality traits or nativity in relation to the language.

The data used in this context comes from various sources, ranging from blog articles, emails, twitter or chat platforms to various social media, opinions and comments from forum users. On the other hand, while the distribution of gender classes is commonly accepted given its binary and discrete character [male, female], it is different from age. Indeed, the granularity of the age classes presents itself differently according to the cases, namely to the unit or according to discrete age groups variously distributed in number and in intervals.

Furthermore, the methods for determining profiles are mostly based on supervised learning [4]. As for the models used, they are commonly based on statistical algorithms. Among which, we cite Bayesian models [5], regression models [6], clustering models and decision trees.

The results obtained are based, depending on the work and methods used, either on lexical (specific vocabulary, variation) or grammatical (syntactic constructions, chords, POS) or even orthographic (quality, punctuation, capital) or stylistic [7] (prototypical expressions, language register, sentiment markers etc.) with differences in pre-processing and segmentation (tokens, Ngram, collocations) [8].

In this paper, author profiling (AP) issue is addressed as aim for this paper. By profiling the author, we mean the estimate of these users' two socio-demographic indicators "Age" and "Gender" based solely on their textual outputs. For example, once a consumer connects to a certain company's website, this organization tries to collect all available data about him or her. After that, the customer receives a request for service according to the age and gender of his or her personal information. Thus, to build an effective categorization, a context-specific choice-making framework based on this data is required. Such a program of choice would encourage businesses to publicize their marketing. To achieving the best possible classification, we follow a range of machine learning strategies like Support Vector Machines (SVM), Random Forest (RF), Multilayer Perceptrons (MP), Decision trees (DT), Naïve bayes (NP) and Long Short-Term Memory(LSTM) as deep learning strategy [9][10]. The adopted corpus is extracted from the dataset PAN-AP-2015, which is in effect accessed from Twitter. Only English language was considered in our work.

The rest of this document is structured as follows. Section 2 is dedicated for the description of the adopted corpus and methodology. Section 3 gives details about the obtained results. Section 4 concentrates on related contributions in the literature. Finally, Section 5 proposes a conclusion for the paper and gives possible extensions for the proposed approach.

2. RELATED WORK AND CORPUS AND METHODOLOGY

2.1. Corpus

The adopted corpus in this work is extracted from the PAN-AP-2015 dataset. The latter is taken from Twitter in the four languages Dutch, Italian, Spanish and English. In our work, only English language is considered. The considered dataset is annotated with ages classes and gender. For age classification, we consider four classes (ie, "from 18 to 24", "from 25 to 34", "from 35 to 49" and "50 and more". The corpus is split into 3 subsets: "training", "early birds" and "test". The distribution of the corpus is illustrated in the following table.

Table 1: Distribution of the adopted corpus

	Training	Early birds	Test
Users	152	42	142
18-24	58	16	56
25-34	60	16	58
35-49	22	6	20
50+	12	4	8
Male	76	21	71
Female	76	21	71

2.2 Architecture of the system

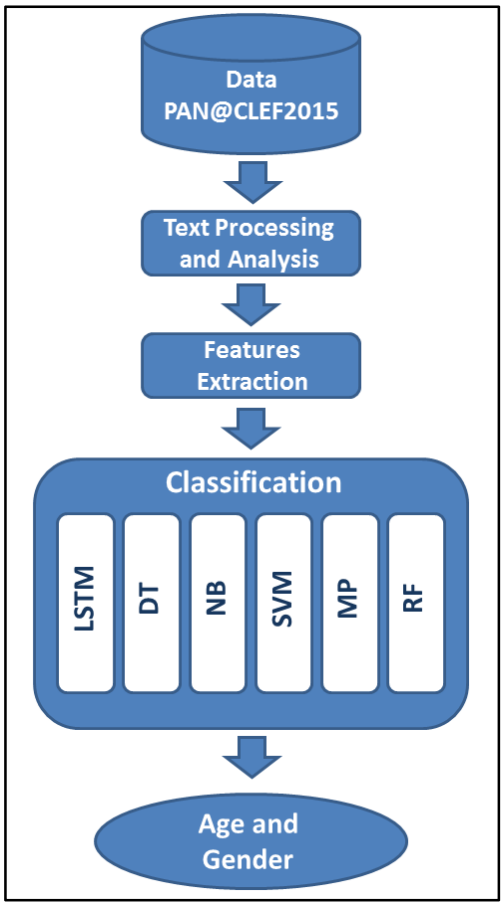

Figure 1: Architecture of the adopted System.

As shown in Figure 1, the architecture of our system is made of the three following blocks:

- Text Processing and Analysis: Some scientists resorted to preprocessing to guarantee accurate prediction of sex and age. For example, "HTML Cleaning" can be used to differentiate between bots and humans. Another strategy is the deletion of Twitter URLs, hashtags and user entries. In addition, conversion of case, and removal of multiple white spaces and invalid characters can be helpful at this stage.

- Features Extraction: The extraction of features is a reduction of dimensionality by which an initial collection of raw data is condensed to more usable groups for processing. A disadvantage of these big data collections is a huge number of variables which take a huge amount of computational capacity to process. Extraction of features is the term for solutions which collect and/or merge parameters into features, effectively decreasing the amount of data to be processed while still accurately and fully representing the initial collection of data. This method is useful for reducing the amount of required resources without missing essential or appropriate information. Extraction of features may also reduce the quantity of redundant data needed for a given analysis. In addition, data reduction and computer's efforts to create variable combinations accelerate the learning progress during the ML process.

- Classification: this step is achieved using one of the following techniques:

    - LSTM: Long Short-Term Memory;
    - DT: Decision Trees;
    - NB: Naïve Bayes;
    - SVM: Support Vector Machines;
    - MP: Multilayer Perceptrons;
    - RF: Random Forest.
    - KNN: k-nearest neighbors

    More details about these six techniques are presented in the next six subsections.

2.3Gender prediction based on LSTM

Recurrent neural networks (RNN) include cycles within the neural graph [11]. The main motivation behind this type of architecture is to be able to manipulate sequences of input vectors, each representing a temporal event, and not just isolated data having no temporal significance.

The advantage of RNNs lies in their ability to consider the past context when processing current information. However, these networks have difficulties in processing relatively long sequences, in particular those containing more than 10 events [12]. Indeed, with cumulative calculations over the long term, the error obtained with the backpropagation of the gradient decreases or, less frequently, increases exponentially with respect to the time scale. These two problems are respectively called ``vanishing gradient'' and ``exploding gradient''[12].

One of the most effective solutions to solve this problem of calculating the gradient is manifested in an extension of the concept of RNN, namely, the Long Short-Term Memory (LSTM) architecture [13].

The peculiarity of LSTM lies in the way in which the hidden state is managed. In the case of simple RNNs, the processing of the recurrence, symbolized by the H function, is ensured by a simple tanh function. In the case of LSTM, this treatment is replaced by a "memory cell". The LSTM cell is characterized by a central node, containing the internal state of the cell, and a number of "doors" divided into 3 categories. These doors make it possible to manage, on the one hand, the keeping in memory of the sequential information (entry and forgetting doors).On the other hand, the role of the internal state in the production of each exit (exit door). By closing the front door, for example, new events are less considered in the information of the cell.

LSTMs have shown their effectiveness in various fields of application. They are currently like the state-of-the-art approach in several tasks dealing with sequential data [14] [15]. Their distribution is manifested especially in the case of fairly long sequences of events [16] [17].

2.4 Gender and age prediction based on machine learning algorithms

A. Decision trees (DT)

Decision trees [18] are architectures that classify the input instances by routing them through conditions posed on the values of the attributes of said instances. In a decision tree, each node represents a specific attribute. A branch emanating from a node represents a condition on the attribute of the same node. Finally, a sheet constitutes a classification decision to be taken following the verification of the conditions posed from the root. Each instance therefore descends level by level, starting from the root, each time crossing the branch validating the condition relating to the value of the upper node.

Regarding the construction of a decision tree, a selection of the attribute that most effectively separates the learning data is done recursively. The first selection thus provides the attribute of the root of the tree accompanied by the conditions (branches) relating to its value. Then, the son node attached to each branch is chosen either as a leaf of the tree, and therefore a class, or as an attribute developing a subtree in the same way as the root node.

From the above, we notice that the selection of the attribute according to which the data will be distributed is a fundamental step. Several methods have been proposed to find the optimal attribute such as information gain [19] and the Gini index [20]. On the other hand, several studies have shown that there is no optimal method [21].

B. Naive bayes(NB)

The naive Bayesian classifier is a generative learning algorithm. Generative models assume that, for a certain class, the sequences (or, in general, the input data) are generated according to a probability law. Naive Bayesian classification is known for its simplicity while being effective. This algorithm is based on the Bayes theorem.

The naive Bayesian classification admits that the existence of an event, for a class, is independent of the existence of other events. When processing sequential data, this assumption is generally not respected. Despite this drawback, and thanks to their simplicity, speed and efficiency, naive Bayesian classifiers have been used in various tasks handling sequential data [22] [23]. [24] even showed the scalability of the naive Bayesian classifier by categorizing millions of opinions on films as part of big data.

C. Support Vector Machines (SVMs)

SVMs (Support Vector Machines) developed by [25] are supervised learning techniques that are among the best performing classification algorithms. SVMs seek to separate two groups of instances (or projections of instances) by a maximum margin hyperplane. Such a hyperplane is considered to be an optimal separator which will have a better ability to generalize and classify the new unknown examples.

Linear SVMs are the simplest form of this algorithm. They are applicable in the case where the data is linearly separable. If the training data are not linearly separable but can be separated by means of a non-linear function, the process of determining the classification function consists in this case of two stages. First, the input vectors are projected into a larger space so that they can be linearly separable. Next, the SVM algorithm is used to find the optimal hyperplane that separates the new data vectors. This hyperplane is therefore defined by a linear function in the new space but with a non-linear function in the original space.

Let **F** be the function of projecting data into the destination space. After this projection, a learning algorithm could only manipulate the data through the dot products in this last space. The kernel functions are special functions which allow to compute the products directly in the original space without going through the projection **F** which we will no longer need to determine. Despite the advantages of kernel functions, their interpretation is difficult, and the user cannot learn from the behavior of the classifiers.

The main motivation behind the use of SVM in the classification of sequential data is the possibility of projecting the sequences into a space of characteristics of larger dimension and of finding there a hyperplane which achieves a maximum margin between the instances of each pair of classes. Therefore, unlike many other non-neural classification algorithms, SVMs do not always need to transform input sequences into feature vectors [26].

D. Multilayer Perceptrons (MP)

Multilayer Perceptrons (MP) are non-looped neural networks whose nodes are organized into three or more levels called "layers". The neighboring layers are completely connected, that is to say, the nodes of each layer are linked to all the nodes of the lower layer and to all those of the upper layer. In contrast, no connection exists between the units of the same layer.

An MP consists of three types of layers, an input layer, a set of hidden layers, and an output layer:

- The input layer is the 1stlayer of the network. Activations of this layer receive the information provided by the input vectors of each instance. This layer therefore has no input connections from other nodes. It is however completely connected to the 1st hidden layer.

- In an MP containing N ($N \geq 1$) hidden layers, each of the N-1 lower hidden layers is completely connected to the one above. The N-th and last hidden layer is completely connected to the output layer.

- The activations of the output layer neurons represent the values of the MP output vector.

MPs are often used in classification tasks [27] [28] but less frequently for processing sequential data [29]. We therefore consider, in the following manuscript, MP as belonging to the category of classical algorithms although these models are distinguished by other functionalities. Thanks to the parametrability of these networks by adjusting the input weights of neurons, MPs have the ability to imitate the behavior of various functions. [30] have even shown that an MP containing a single hidden layer and enough neurons with nonlinear activations can be close to any continuous function. For this reason, MPs are considered to be universal functions. In addition, these networks can be used to generate more robust representations of the data.

E. Random Forest (RF)

Random Forest [331-32] (or random forest decision) is a way of thinking about grouping, regression, and other tasks together. At the time of training, this approach operates by constructing a number of decision trees and generating the class that is the class mode (classification) or the median

prediction (regression) of the trees. Random forests fix the practice of "overcrowding" decision trees with respect to their collection of training.

Tin Kam Ho [34] created the first algorithm for random decision forests utilizing the "random subspace" technique, that is a way to enforce the "stochastic discrimination" methodology to the grouping suggested by Eugene Kleinberg.

F. k-Nearest Neighbors (KNN)

It is a type of classification algorithm. It predicts a new data point by finding the nearest training set depending on fixed number k (distance function) to that point and assign it to the training set. It is the easiest machine learning algorithm to do and understand. It gives good performance without a lot of adjustment. However, it is slow and cannot handle many features [17]

2.5 Related work based on deep learning techniques

There are several studies focused on the linguistic characteristics. The authors of [36], for example, tried to illustrate the role of personal phrases (phrases with a first-person pronoun) for AP purposes. Two new ideas are introduced for that reason, namely: a "feature selection process" and a "time weighting scheme'.' Those two measures are based on a new metric called "Personal Expression Intensity" (PEI) that measured the amount of information transmitted over a given period. The suggested method was conducted empirically to forecast age and gender for media viewers who depend on 6 separate samples of data (in English).

Some other research proceeded to n-grams to determine the authors' profile. The authors of [37], for example, established new SMS (short text) profiling data in both Roman Urdu and English. Two forms of features have been used: (1) stylistic features like lexical terms and paragraph and verbs counts; (2) material features centered on n-grams ranging between 2 and 10. Empirical research showed excellent results, especially for the richness of vocabulary.

Several studies in the literature adopted certain forms of features in this same line of thinking in order to supplement the thematic features. For example, the authors of [38] conducted a thorough study of the importance of "distributional term representations" (DTRs) to tackle the issue of AP in social media. They provided a supervised system for the AP using DTRs to do so. In the AP mission, they assessed the "text occurrence representation' (DOR) and the "word co-occurrence representation" (TCOR). Using the suggested AP system, they provided a comparative study of several representations on the distribution. Researchers have measured their results against the findings of classic bag-of-words and success approaches focused on topics.

Methods for DL were recently hired. For example, the authors of [39] used two DL methods to solve the problem of sex and age recognition for the case of Lithuanian neural word embeddings. The first is "Long Short-Term Memory' (LSTM) and the second is "Convolutionary Neural Network' (CNN).' They analyzed the influence of the training database on the efficacy of the AP by the scale. It has been shown that the LSTM approach is more efficient for limited datasets, while the CNN approach is shown to be mostly efficient for bigger data sets. The researchers contrasted the techniques used to classic ML techniques (especially "Support Vector Machine and Multinomial Naive Bayes'.' This analysis shows that for the considered AP problem, the DL strategies used are less effective than the classical ML strategies.

The authors of [40] employed another form of deep learning. The researchers used a hybrid RNN and CNN method for gender identification in both the English and Spanish languages in this research. The model suggested for this is focused on n-grams. In both RNN and CNN models 1024 neurons were used for activation with ReLU. The researchers analyzed the PAN2019 corpus to assess the existing framework.

2.6 Evaluation metrics

Since our training dataset is roughly unbalanced, and as we mentioned above, the use of different performance measures can be a useful solution to overcome the problem of imbalanced data in order to evaluate the classification model. Thus, to evaluate the performance of our method, we employed three metrics commonly used in data mining evaluation: accuracy, precision and recall [9]. These can be simply calculated from a confusion matrix with the equations (4), (5) and (6), as shown at Table 2

Table 2: Confusion matrix: displays the number of samples that were classified correctly (TP and TN) and falsely (FP and FN).

Confusion matrix		**Predicted class**	
		Class A	Class b
True class	Class A	True Positives (TP)	False Negatives (FN)
	Class b	False Positives (FP)	True Negatives (TN)

Where True Positives (TP) and True Negatives (TN) are respectively the number of positive examples and negative examples, labelled as such. Conversely, False Positives (FP) and False Negatives (FN) are the number of positive examples and negative examples, which are falsely labelled.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN))} \quad (1)$$

$$\text{Precision} = \frac{TP}{(TP + FP)} \quad (5)$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \quad (6)$$

3. EXPERIMENTATION AND RESULTS

In this experiment, we test some knowns algorithms for gender and age detection. We note that best results for age and best results for gender are not obtained by the same model. For the gender prediction (Table 3) we tried several classification models, the best result is obtained using the algorithm of the family of Bayesian mutinomial NB models with an accuracy of 0.96 against 0.66 for decision trees (J48) and 0.61 for Naive Bayes. Also, in terms of precision and recall we obtained the best score with Multinomial NB (0.7 for the precision, 0.56 for the recall and 0.59 0.64 for the F-score). Note that the Baseline is equal to 0.5 because it is to predict two classes only. The Baseline outperform the scores obtained by the KNN and the naïve Bayes algorithms.

In a new experiment, we explored the deep learning track with LSTM using 8 epochs, but we had average results because the base is too sma55ll for deep learning. The results are disappointing (only 0.51 for the Accuracy) because the base is too small for deep learning.

Table 3: Result of Model for Gender

Model for Gender	Precision	Recall	Accuracy	F-score
Decision trees	0.7	054	066	0.6
Multinomial NB	0.75	0.56	0.69	0.64
Multilayer Perceptron	0.7	055	0.61	0.59
LSTM			0.51	
Naïve Bayes	0.4	0.4	0.3	0.4
KNN	0.42	0.42	0.7	0.42
Baseline	**0.5**	**0.5**	**0.5**	**0.5**

Unlike gender (age in Table 4) and by applying the rule, there is not one model that always succeeds, we obtained the best score by applying neural networks. Indeed, we obtained an accuracy of 0.63 with Neural network against 0.58 and 0.62 for multinomial NB and Decision trees. However, the best recall is obtained with Mutinomial NB. Note that the base line is ¼ since it is a question of predicting 4 classes. The KNN classifier gives disappointing results on text data. The accuracy is only 0.5.

Table 4: Result of Model for Age

Model for Age	Precision	Recall	Accuracy	F-score
Decision Trees	0.61	0.70	0.58	0.65
Multinomial NB	0.60	0.88	0.62	0.72
Multilayer Perceptron	0.59	0.78	0.63	0.67
LSTM			054	
Naïve Bayes	0.61	0.61	0.5	0.61
KNN	0.3	0.3	0.4	0.3
Baseline	**0.25**	**0.25**	**0.25**	**0.25**

We conclude that for gender prediction the best results are obtained by multinomial NB. However, for the age the best performance is obtained by the multilayer perceptron (neural network). For the two dimensions deep learning fail to obtain good results.

4. CONCLUSION

Within this paper we discussed the problem of author profiling (AP) using machine learning [41][43] and deep learning [43] [44] techniques. The adopted corpus was taken from the PAN-AP-2015 dataset, which is directly accessed from Facebook. During our research only English was taken into consideration. To achieve the best possible classification, we adopted a number of machine learning techniques Decision trees, Naïve bayes[46], Support Vector Machines [47], Multilayer Perceptrons, Random Forest and deep learning (Long Short-Term Memory "LSTM") [48].The results show that each data set (age or gender) have different results when machine learning techniques were applied. This is related to the power of each used technique as mentioned in [9][48]. Deep learning techniques proved that they will be helpful when the data set are large as mentioned in [49].

Among the possible future directions to extend our work we cite:

- Use a big data corpus to obtain better results with deep learning.
- Test if our method is multilingual on the same corpus since there are a sub-corpus in Italian and Spanish.
- Test if our method is reusable on new corpora, notably the pan@clef2019.
- Adopt a set of appropriate model-based testing techniques [50] in order to validate to proposed approach.

REFERENCES

1. Ali H. Al-Badi, Michelle, O. Okam, Roobaea Al Roobaea and Pam J. Mayhew (2013), "**Improving Usability of Social Networking Systems: A Case Study of**

LinkedIn," *Journal of Internet Social Networking & Virtual Communities*, Vol. 2013 (2013), Article ID 889433, DOI: 10.5171/2013.889433

2. Nadeau D. &Sekine S. (2007). **A survey of named entity recognition and classification**. *LingvisticaeInvestigationes*, 30(1), 3–26.
3. Rangel F., Stamatatos E., Moshekoppel M., Inchesg. & Rosso P. (2013). **Overview of the author profiling task at pan 2013**. *In CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, p. 352–365: CELCT.
4. Rangel F., Rosso P., Potthsst M., Stein B. &Daelemanss W. (2015). **Overview of the 3rd Author Profiling Task at PAN 2015**. *In CLEF 2015 Evaluation Labs and Workshop Working Notes Papers* (pp. 1-8).
5. Chen J., Huang H., Tian S. & QU Y. (2009). Feature selection for text classification with naïve bayes. Expert Systems with Applications, 36(3), 5432–5435. https://doi.org/10.1016/j.eswa.2008.06.054
6. Nguyen D., Smith N. A. & Rose C. P. (2011). **Author age prediction from text using linear regression**. *In Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH'11, p. 115–123, Stroudsburg, PA, USA: Association for Computational Linguistics.
7. Weren E. R., Kauer A. U., Mizusaki L., Moreira V. P., DE oliveira J. P. M. & Wives L. K. (2014). **Examining multiple features for author profiling**. *Journal of Information and Data Management*, 5(3), 266.
8. Santosh K., Bansal R., Shekhar M. & Varma V. (2013). **Author profiling: Predicting age and gender from blogs—notebook for pan at clef 2013**.
9. Alsufyani, A., Alroobaea, R. & Ahmed, K.A.(2019). **Detection of single-trial EEG of the neural correlates of familiar faces recognition using machine-learning algorithms**. *International Journal of Advanced Trends in Computer Science and Engineering*. [Online] 8 (6). Available from: doi:10.30534/ijatcse/2019/28862019.
10. S Mechti, Alroobaea, R., M., Krichen., Rubaiee, S. and Ahmed, A., 2020. **Deep Learning Model for Identifying the Arabic Language Learners based on Gated Recurrent unit Network.** *International Journal of Advanced Computer Science and Applications (IJACSA)*, 11(5).
11. J. L. Elman, 1990. **Finding structure in time. Cognitive science** 14(2), 179–211.
12. S. Hochreiter, Y. Bengio, P. Frasconi, & J. Schmidhuber, 2001. **"Gradient flow in recurrent nets: the difficulty of learning long-term dependencies**.
13. S. Hochreiter & J. Schmidhuber, 1997. Long shortterm memory. Neural computation 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
14. K. Yao, B. Peng, Y. Zhang, D. Yu, G. Zweig, & Y. Shi, 2014. **Spoken language understanding using long short-term memory neural networks**. *Dans les actes de Spoken Language Technology Workshop (SLT)*, 2014 IEEE, 189–194. IEEE. https://doi.org/10.1109/SLT.2014.7078572
15. T. Fischer & C. Krauß, 2017. **Deep learning with long short-term memory networks for financial market predictions.** *European Journal of Operational Research*, 270(2), pp.654-669.
16. X. Ma, Z. Tao, Y. Wang, H. Yu, & Y. Wang, 2015. **Long short-term memory neural network for traffic speed prediction using remote microwave sensor data**. *Transportation Research Part C: Emerging Technologies* 54, 187–197. https://doi.org/10.1016/j.trc.2015.03.014
17. D. Li & J. Qian, 2016. **Text sentiment analysis based on long shortterm memory**. *Dans les actes de International Conference on Computer Communication and the Internet (ICCCI)*, 471–475. IEEE.
18. J. R. Quinlan, 1986. Induction of decision trees. Machine learning. *Kluwer Academic Publishers, Boston - Manufactured in The Netherlands*, 1(1), 81–106.
19. E. B. Hunt, J. Marin, & P. J. Stone, 1966. **Experiments in induction**.*Academic Press, New York*.
20. L. Breiman, J. Friedman, C. J. Stone, & R. A. Olshen, 1984. *Classification and regression trees.CRC press.*
21. S. K. Murthy, 1998. **Automatic construction of decision trees from data: A multi-disciplinary survey**. *Data mining and knowledge discovery* 2(4), 345–389. https://doi.org/10.1023/A:1009744630224
22. B. Y. M. Cheng, J. G. Carbonell, & J. Klein-Seetharaman, 2005. **Protein classification based on text document classification techniques**. *Proteins: Structure, Function, and Bioinformatics* 58(4), 955–970.
23. Z. Muda, W. Yassin, M. Sulaiman, & N. Udzir, 2016. **K-means clustering and naive bayes classification for intrusion detection**. *Journal of IT in Asia* 4(1), 13–25.
24. B. Liu, E. Blasch, Y. Chen, D. Shen, & G. Chen, 2013. **Scalable sentiment classification for big data analysis using naive bayes classifier**. *Dans les actes de International Conference on Big Data*, 99–104. IEEE. https://doi.org/10.1109/BigData.2013.6691740
25. V. Vapnik, 1999. **The Nature of Statistical Learning Theory.***Springer Science & Business Media.*
26. Z. Xing, J. Pei, & E. Keogh, 2010. **A brief survey on sequence classification.** *ACM Sigkdd Explorations Newsletter* 12(1), 40–48.
27. B. Chaudhuri & U. Bhattacharya, 2000. **Efficient training and improved performance of multilayer perceptron in pattern classification**. *Neurocomputing* 34(1), 11–27. https://doi.org/10.1016/S0925-2312(00)00305-2
28. S. L. Phung, A. Bouzerdoum, & D. Chai, 2005. **Skin segmentation using color pixel classification: analysis and comparison.***IEEE transactions on pattern analysis and machine intelligence* 27(1), 148–154.
29. Y.-P. Lin, C.-H. Wang, T.-L. Wu, S.-K. Jeng, & J.-H. Chen, 2007. **Multilayer perceptron for EEG signal classification during listening to emotional music**. *Dans les actes de TENCON Region 10 Conference*, 1–3. IEEE.
30. K. Hornik, M. Stinchcombe, & H. White, 1989. **Multilayer feedforward networks are universal approximators**. *Neural networks* 2(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8
31. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, 2003, **Feuston BP. Random forest: a classification and**

regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences.* Nov 24;43(6):1947-58.

32. Ham J, Chen Y, Crawford MM, Ghosh J.2005 **Investigation of the random forest framework for classification of hyperspectral data**. *IEEE Transactions on Geoscience and Remote Sensing*Feb 22;43(3):492-501. https://doi.org/10.1109/TGRS.2004.842481
33. Rogers J, Gunn S**.**2005**, Identifying feature relevance using a random forest**. *InInternational Statistical and Optimization Perspectives* Workshop" Subspace, Latent Structure and Feature Selection" Feb 23 (pp. 173-184). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11752790_12
34. Ho TK.1995,**Random decision forests.InProceedings of 3rd international conference on document analysis and recognition**, (Vol. 1, pp. 278-282). IEEE.
35. Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A., 1984. **Classification and regression trees**. *CRC press*.
36. Ortega-Mendoza RM, López-Monroy AP, Franco-Arcega A, Montes-y-Gómez M.2018,**Emphasizing personal informationfor Author Profiling: New approaches for term selection and weighting**.*Knowl. -Based Sys*t; 145: 169–181.
37. Fatima M, Anwar S, Naveed A, Arshad W. 2018, **Multilingual SMS-based author profiling: Data and methods**.*Natural LanguageEngineering*; 24(5): 695–724.
38. Carmona MÁÁ, Villatoro-Tello E, Montes-y-Gómez M, Pineda LV. 2019, **A comparative analysis of distributional term representations for author profiling in social media**.*Journal of Intelligent and Fuzzy Systems;* 36(5): 4857–4868. https://doi.org/10.3233/JIFS-179033
39. Kapociute-Dzikiene J, Damasevicius R. Lithuanian,2018, **Lithuanian author profiling with the deep learning**. *In 2018 Federated Conference on Computer Science and Information Systems (FedCSIS*) (pp. 169-172). IEEE
40. Dias RFS, Paraboni I.2019,**Combined CNN+RNN Bot and Gender Profiling**. In: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12.
41. Gadade, H. D., andKirange, D.K., 2020. **Machine Learning Approach towards Tomato Leaf Disease Classification**. *International Journal of Advanced Trends in Computer Science and Engineering* 9(1):490-495. https://doi.org/10.30534/ijatcse/2020/67912020
42. Buladaco, M. V. M., Buladaco, J. S., andCantero, L.M., 2020. **Sentiments Analysis on Public Land Transport Infrastructure in Davao Region using Machine Learning Algorithms**.*International Journal of Advanced Trends in Computer Science and Engineering* 9(1):685-690. https://doi.org/10.30534/ijatcse/2020/97912020
43. Mercara, J. L. D.,Delima, A. J. P., andVilchez, R. N., 2020. **Prediction of Employees' Lateness Determinants using Machine Learning Algorithms**. *International Journal of Advanced Trends in Computer Science and Engineering* 9(1):779-783. https://doi.org/10.30534/ijatcse/2020/111912020
44. Ng, C. and Chua, A., 2020. **Training of a deep learning algorithm for quadcopter gesture recognition**. *International Journal of Advanced Trends in Computer Science and Engineering* 9(1):211-216. https://doi.org/10.30534/ijatcse/2020/32912020
45. Keerthana, R. and Chooralil, V. S., 2020. **Forecasting of the Air Pollution Based on Meteorological Data and Air Pollutants using Deep Learning: A Novel Review**. *International Journal of Advanced Trends in Computer Science and Engineering* 9(1):194-200. https://doi.org/10.30534/ijatcse/2020/115912020
46. Denila, P. G., Delima A. J. P., and Vilchez, R. N., 2020. **Analysis of IT Graduates Employment Alignment Using C4.5 and Naïve Bayes Algorithm**. *International Journal of Advanced Trends in Computer Science and Engineering* 9(1):745-752. https://doi.org/10.30534/ijatcse/2020/106912020
47. AL-Shatnawi, A., Al-Saqqar, F. and Al-Smadi, H., 2020. **A hybrid Feature Selection Approach for Arabic Handwritten Text Based on Genetic Algorithm and Support Vector Machine**. *International Journal of Advanced Trends in Computer Science and Engineering* 9(1):813-819. https://doi.org/10.30534/ijatcse/2020/117912020
48. Rhemimet, A., Raghay, S., Bencharef, O. and Chihab, Y., 2020. **Long Short-Term Memory Recurrent Neural Network Architectures for Prediction of HIV-1 Protease Cleavage Sites**. *International Journal of Advanced Trends in Computer Science and Engineering* 9(1):194-200. https://doi.org/10.30534/ijatcse/2020/29912020
49. Mechti S, Jaoua M, Faiz R, Bouhamed H, Belguith LH. **Author Profiling: Age Prediction Based on Advanced BayesianNetworks**.*Research in Computing Science2016*; 110: 129–137.
50. Krichen, M. and Alroobaea, R., 2019, **Towards Optimizing the Placement of Security Testing Components for Internet of Things Architectures**. *In 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA),* (pp. 1-2). IEEE https://doi.org/10.1109/AICCSA47632.2019.9035301