IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 February 2023, accepted 17 March 2023, date of publication 24 March 2023, date of current version 29 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3261339

RESEARCH ARTICLE

Words Similarities on Personalities: A Language-Based Generalization Approach for Personality Factors Recognition

ADRIANO MADUREIRA DOS SANTOS[1], FLÁVIO RAFAEL TRINDADE MOURA[1], LYANH VINICIOS LOPES PINTO[1], ANDRÉ VINICIUS NEVES ALVES[1], KARLA FIGUEIREDO[2], FERNANDO AUGUSTO RIBEIRO COSTA[3], AND MARCOS CÉSAR DA ROCHA SERUFFO[1]

[1]Institute of Technology (ITEC), Federal University of Pará (UFPA), Belém 66075-110, Brazil
[2]Computational Intelligence and Robotics Laboratory (LIRA), Rio de Janeiro State University (UERJ), Rio de Janeiro 20950-000, Brazil
[3]Center for High Amazon Studies (PDTU/NAEA), Federal University of Pará (UFPA), Belém 66075-110, Brazil

Corresponding author: Adriano Madureira dos Santos (adrianomadureira1@gmail.com)

This study was supported in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) under Finance 001, in part by the National Counsel of Technological and Scientific Development (CNPq), and in part by the Federal University of Pará (UFPA) by the Pró-Reitoria de Pesquisa e Pós-Graduação (PROPESP).

ABSTRACT The evaluation of personality traits allows the study of human behavior in different environments, but it is not a trivial task. In this sense, the Five-Factor Model (FFM) allows, in a global way, the assessment of personality traits of individuals using textual data. However, there is a scarcity of lexical resources for languages other than English, which generated the main research question of this work: ''Can models trained to predict FFM personality traits using English textual data show satisfactory results when applied to textual data in other languages?''. Therefore, this work aims to answer: (i) Whether Word Embeddings techniques could be used to solve low resources languages problems in FFM personality traits prediction; and (ii) Whether is feasible to train a traditional Machine Learning algorithm with English language textual data and evaluate its performance with Brazilian Portuguese language textual data for FFM personality traits prediction. Thus, the work aims to present an approach in which the models can be used to learn the highest level of abstraction. As results, was observed that the difference in performance between the models trained for personality recognition in English is minimal when used to predict FFM personality traits in Brazilian Portuguese texts. In this task, the Stochastic Gradient Descent model presented the best average results among the FFM personality traits of the models analyzed.

INDEX TERMS Machine learning, natural language processing, online social networking, technology social factors, knowledge transfer.

I. INTRODUCTION

Knowledge of individuals' personalities makes it possible to identify patterns and behaviors that may be more appropriate for specific contexts. Among the main applications that make use of this information are recruitment and recommendation [1] and psychological and behavioral profiling systems for a given activity [2], [3]. In this sense, personality detection is an innovative field capable of tailoring services to individual interests and identifying anomalous behavioral traits, presenting useful applications for society [4], [5]. The study of personalities started from the analysis of words and their correlations with the individuals' behaviors.

In this context, the lexical hypothesis, which suggests that fundamental human personality traits have, over time, been encoded in language, has been widely used to study the structure of personality traits in various cultural and linguistic settings [6]. This hypothesis is commonly defined by two postulates, where it is stated that personality traits that are important to a group of people eventually become part of

The associate editor coordinating the review of this manuscript and approving it for publication was Agostino Forestiero.

that group's language, and that the most important personality traits tend to be encoded as single words in language [7].

However, the assessment of personality traits is not a trivial task. The difficulties of manual data labeling of personality traits are susceptible to human subjectivity [8]. Due to this, the Five Factor Model (FFM), also known as the Big Five, is currently considered the most successful effort to assess the personality traits of individuals in a global way [9]. The FFM was initially intended for personality classification through lexical patterns. Other personality theories such as HEXACO[1] and Myers-Briggs,[2] are also frequently used in scientific circles. However, comparatively, the Big Five brings together more general personality traits than the others, being more frequently used in scientific research to assess personality traits [10].

According to [11], [12], [13], and [14] the five major factors of the FFM are: Openness to experience (O), which sometimes is called intelligence; Conscientiousness (C); Extroversion (E); Agreeableness (A); Neuroticism (N), which sometimes is called the opposite of emotional stability. Openness to experience people are broad rather than narrow in their interests and prefer novelty to routine. Conscientious people are task-oriented, rather than distracted or disorganized. Highly extroverted people tend to be assertive as well, rather than peaceful and reserved. Agreeableness people tend to be cooperative and polite, rather than hostile and rude. Finally, people with high neuroticism are more likely to experience negative emotions than emotionally resilient people. The levels of the five personality factors of an individual can be obtained through FFM questionnaires [15], [16].

Personality traits are reflected in many different environments, including on online platforms such as Online Social Networks (OSN) [17], [18]. Specifically, these environments had an intensification in its use during the last years, providing significant changes in the dynamics of social interaction, in addition to a greater amount of access due to the participation of individuals in themes that have repercussions in society [19], [20]. This participation also allows the dissemination of contents that can carry personality peculiarities among different contexts and impact the environment users' lives, leading individuals to act together, whether these actions are beneficial to the environment or not. For example, [21] and [22] found that digital agression is more practiced by high extroverted and low conscientious people. Similarly, the research conduct by [23] found that the digital agression is more accentuated to low conscientious people.

In this sense, the investigation of OSN is fundamental to understanding the structure of personality traits, allowing an understanding of the dynamics of social interaction and social problem solving, including the social behavior [24]. Furthermore, the constant use of the OSN by its users provides an increasing amount of textual content available, enabling the use of techniques and methods for identifying patterns through textual data. These patterns can be reflected in the personality traits and behaviors of individuals, as well as enabling interpretation about the individual intentions of each user in the environment [25].

Although the task of automatic personality trait recognition can provide improvements in analyses performed on online environments using textual data; there is a low amount of lexical resources available for languages other than English [26]. The difficulties in finding annotated personality datasets, as well as the appropriate tools to extract attributes from textual data, such as Linguistic Inquiry and Word Count[3] (LIWC) [27], are recurrent aspects for the other languages [28]. This manifests difficulties in applying techniques designed to recognize personality traits in languages that differ from English.

Hence, the main research question was formulated: ''Can models trained to predict FFM personality traits using English textual data show satisfactory results when applied to textual data in other languages?''. In this sense, the investigation of solutions to this problem can contribute to the scientific community in the study of the automatic FFM personality traits recognition in languages with a low amount of textual resources, besides opening a discussion about the similarities that exist between the semantics of terms written in different languages.

Thus, this paper aims to evaluate the feasibility of train a traditional Machine Learning algorithms with English textual data and evaluate its performance with Brazilian Portuguese data for prediction of FFM personality traits. To prepare the textual data for the use of the models, Natural Language Processing techniques capable of constructing word vectors considering semantic and syntactic similarity measures will be used. The intention is to present an approach in which the models can be used to learn the highest level of abstraction: the semantics that can be mapped through sets of different words among languages.

Still, the method proposed in this work, which uses traditional machine learning models combined with word embedding techniques of different languages, presents close results to works that use only one language for the task, as well as surpasses approaches that do not depend on extracted resources of the language. In addition, the method consumes less computational resources, requiring less training time, and is also capable of providing robust results when compared to other models. Therefore, the method developed is innovative and differs from the other methods found.

The main contributions of the paper are described below:

(i) Word Embeddings techniques could be used to solve low resources languages problems in FFM personality traits prediction;
(ii) Train traditional Machine Learning algorithm with features extracted from English textual data could be

[1]https://hexaco.org/scaledescriptions
[2]https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/
[3]LIWC is a text analysis tool that allows to perform psycholinguistic attribute extraction.

a solution to predict FFM personality traits through Brazilian Portuguese textual data.

The remainder of this paper is divided into 5 more sections. Section II will present the related works for the proposed method. Then, section III will deal with the methodology developed in the paper. The results obtained will be described in section IV. The discussion and comparisons with related works of the proposed model are presented in section V. Finally, the final considerations are presented in section VI.

II. RELATED WORKS

Automatic FFM personality traits recognition using textual data is not a new task. Previously, the task has involved the use of traditional Machine Learning models coupled with tools for extracting psycholinguistic attributes from texts, such as LIWC. In [29], the use of Decision Tree (DT), K-Nearest Neighbors (KNN), Naive Bayes (NB), Ripper, AdaBoost (AB) and Support Vector Machine (SVM) models in regression was performed for personality recognition using the LIWC and Medical Research Council (MRC) Psycholinguistic tools for English language texts from the Essays dataset.

Similarly, these tools were used in [30] for attribute extraction from English-language texts from Twitter, aiming to perform FFM traits recognition with the Gaussian Process and ZeroR regression models. In contrast, [31] performed an open vocabulary approach in FFM traits classification using Differential Language Analysis (DLA). The intent was to find English language features that differ from psycholinguistic attributes, showing that attributes not captured through LIWC can also show significant results in performance.

For FFM personality traits prediction, [32] used attributes extracted from LIWC, Social Network Analysis (SNA) and Structured Programming for Linguistic Cue Extraction (SPLICE) on English language texts from a manually collected dataset and myPersonality. The data was used to train the SVM, Gradient Boosting, Logistic Regression, and XGBoost classification models.

In addition to psycholinguistic attribute extraction techniques, in recent years, Word Embeddings techniques have gained notoriety for the task. These techniques allow to perform word vectorization, making terms with similar meanings have a similar representation [33]. In [34] an exploration of the capabilities of Convolutional Neural Network (CNN) models in classifying FFM traits using Word Embeddings and the Facebook myPersonality dataset was performed.

Additionally, in [35] the recognition of FFM personality traits in short texts through the proposed C2W2S4PT model using Word Embeddings applied to the PAN 2015 dataset was performed. Also, at [36] FFM personality detection occurs from OSN clusters using Network Representation Learning applied to texts. The paper proposed the AdaWalk model, which also uses Word Embeddings and was applied to the Essays dataset. Further, [37] performed FFM personality traits prediction using Facebook statuses using Fully-Connected (FC), CNN and Recurrent Neural Networks (RNN) neural network architectures. In the paper, Word Embeddings was applied to a Facebook dataset and myPersonality.

Also, more recent Language Modeling techniques were used in order to obtain contextual word vectors, more specifically, with the use of Bidirectional Encoder Representations from Transformers (BERT). In [38] transliteration information was extracted from a dataset of YouTube personalities. The intent of the work was to perform a FFM traits prediction approach in which the Word2Vec, GloVe and BERT Embeddings methods could be used to extract word vectors from the transliterations. The extracted vectors were used to train SVM models for classification and also for regression.

Moreover, in [39] a model for personality detection was proposed, combining BERT Embeddings to encoding text vectors with a neural network. The method combined semantic and emotional attributes extracted from Essays datasets in order to perform the training of the CNN, Gated Recurrent Unit (GRU) and Long Short Term Memory (LSTM) neural networks. Besides, in [40] a FFM traits prediction approach was performed in which pre-trained BERT and Robustly Optimized BERT Pretraining Approach (RoBERTa) Contextual Embeddings were used to extract vector representations from the Essays and FriendPersona datasets, the latter constructed by the authors. The vectors formed were used to train the Hierarchical CNN (HCNN), Attention-Based CNN (ABCNN) and Attention-based Bidirectional LSTM (ABLSTM), Hierarchical Attention Network (HAN), BERT and RoBERTa.

The majority of the works mentioned above used feature extraction techniques, such as LIWC and Word Embeddings, to deal with the prediction of FFM traits only focused on English language textual data, which reflects on the scarcity of works focused on personality prediction in other languages. Another example can be verified in [41], in which a study was carried out about the cross-domain intersection between Facebook and Twitter texts in the recognition of FFM traits. The intent of the work was to use the FastText Word Embeddings technique to vectorize English language texts from both social networks. The Facebook data was used to train the SVR, LASSO Regression and Linear Regression (LR) models, while the Twitter data was intended to evaluate the performance of the models.

In [42] a CNN with vectors from the Word2Vec method of Word Embeddings was used as an attribute extractor of English language texts from the Essays dataset. The extracted vectors were used to train the Multilayer Perceptron (MLP) and SVM classification models. Likewise, [43] proposed the FFM traits detection model, which combines a Bidirectional LSTM (Bi-LSTM) neural network architecture and a CNN, named 2CLSTM. In the work, the word vectors coming from the GloVe method of Word Embeddings were used for the extraction of attributes from English language texts, i.e. coming from the Essays and Youtube Personality Datasets. The intent was to use the extracted attributes to train the classification models.

Although few efforts have been made for the prediction of FFM traits in languages that differ from English, Word Embeddings techniques have also been used in order to propose solutions to low-resource problems. For example, in [34] a bilingual model for the classification of FFM traits was proposed, relying on the training of a Word Embeddings model on corpus of English and Chinese language textual data. The intent of the work was to address the problem of data scarcity for personality prediction in the Chinese language. In [44], a personality alignment method was built, named GlobalTrait, in which a multilingual setup was used for personality recognition in English, Spanish, German and Italian languages. In the paper, the Word Embeddings model was trained on a corpus with texts from the four distinct languages to extract attributes from the PAN 2015 dataset.

Other simpler methods differents from Word Embeddings, such as LIWC, were also used for the prediction of FFM traits. For example, in [45], where features were extracted from the PAN 2015 dataset using LIWC, FFM traits prediction occurred from training Stochastic Gradient Descent (SGD) models for classification and Ensemble of Regressor Chains Corrected (ERCC) models for regression. Finally, in [46] the task was performed from sentimental, emotional and social attributes extracted using the National Research Council (NRC) Emotion Lexicon on Indonesian language texts obtained from Twitter. The extracted attributes were used to train NB, KNN and SVM models.

Although the aforementioned research provides satisfactory results, little support was observed for the identification and analysis of textual patterns in Brazilian Portuguese language data. This includes a lack of resources in attribute extraction tools, such as LIWC, as well as public textual data for the recognition of FFM traits in the language. Furthermore, the scientific literature lacks approaches where a Machine Learning model is trained with textual data in one language and evaluated with data from another language, considering the prediction of FFM traits. This shows that a smaller amount of research evaluates the generalization capability of models across texts in different languages. There is even a lack of analysis of the similarities that can be captured between terms present in texts originating from different cultures. Finally, it was also possible to verify that there is a lack of literature on the task for low-resource languages, such as Brazilian Portuguese.

The differential of this proposal is to evaluate the feasibility of forming a model with textual data in English and assess whether its FFM traits predictions made on textual data in Brazilian Portuguese are satisfactory. For this purpose, similarities present in the textual data of both languages are also explored. The model performance results are compared with two different FFM personality traits recognition approaches, one that uses textual data in a single language, i.e. English, and another that is not language dependent for the task.

Thus, the main contributions of this work are divided into two parts. First, Word Embeddings can be used as a way to handle problems of low amount of textual features on low-resource languages, such as Brazilian Portuguese, for FFM traits prediction. Second, the proposed method satisfactory performs the FFM traits recognition on features extracted from Brazilian Portuguese language texts through a model trained with attributes extracted from English language texts.

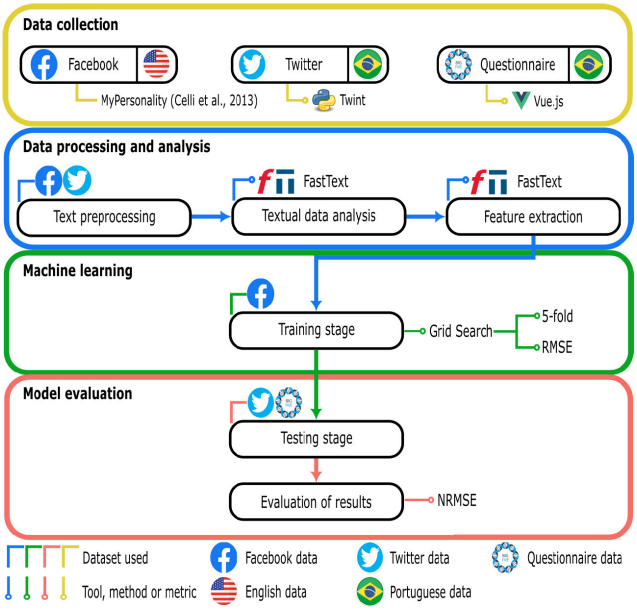


FIGURE 1. Schema with the methodology adopted in the work.

III. METHODOLOGY

The general methodology of this work is presented in Figure 1. The procedures of Data Preprocessing, training and testing Machine Learning models in Computational Intelligence and the Performance Evaluation of the chosen model include the phases and steps necessary for the development of the proposal.

A. DATA COLLECTION

To perform the automatic recognition of personality patterns, was initially collected data available in the literature. The dataset selected for training is a subset of the database *myPersonality*,[4] project developed by David Stillwell and Michal Kosinski, being initially made available by [47]. The *myPersonality* dataset is no longer publicly available, as it has been removed from its authors' platform. For more information about the textual units available in this dataset is displayed in Table 1.

This dataset gathered personality patterns obtained through self-assessment by applying personality questionnaires to volunteers of the MyPersonality project. Each user profile present in the dataset may have more than one English-language post on the *Facebook* platform, and therefore more than one textual content linked to their personality traits.

[4]http://mypersonality.org

TABLE 1. Information associated with the myPersonality dataset.

Metric	Count
Users	250
Publications	9.917
Terms	146.688
Single Terms	15.038

TABLE 2. Information associated with the Twitter dataset.

Metric	Count
Users	106
Publications	4.681
Terms	60.576
Single Terms	17.352

After obtaining the myPersonality dataset, the procedure for collecting personality data was subsequently carried out in order to link it to textual data in Brazilian Portuguese. The purpose of building a dataset for the Brazilian Portuguese language was to evaluate the generalization capability of the model chosen for the automatic recognition of personalities in the testing stage.

In this context, the personality data for the link with textual data in Brazilian Portuguese language, were collected through a website created based on a template made available by Enge et al. (2020), modifying the source code with the Framework Vuejs.[5] This website was made available at <https://tcc-delta.vercel.app/pt>, allowing the research volunteers to enter their Twitter username and answer the IPIP-NEO-120 questionnaire [49].

The personality questionnaire applied to the Brazilian volunteers of the research allowed obtaining the personality scores of the 5 factors and their respective 6 facets, between 24 and 120 points, as well as the detailing of each facet and its implications on the volunteers' behaviors. After completing the questionnaire, the information associated with the volunteers' personality traits is sent to a datasets.

The research volunteers allowed the collection of their textual data in Brazilian Portuguese language published in the social network Twitter. In this context, the Twint library of the Python programming language was used to collect the publications and subsequently link the textual content in Brazilian Portuguese language of each Twitter user profile to its FFM personality traits. Table 2 presents the general information of the textual dataset obtained through the Twitter social network.

The data regarding the FFM personality traits obtained through the questionnaires are converted to the [1], [5] scale for each of the major factors. This procedure is done to ensure compatibility between the patterns of personalities associated with users present in different datasets (i.e. Facebook in English and Twitter in Brazilian Portuguese).

[5]https://vuejs.org

TABLE 3. Information associated with the myPersonality dataset post preprocessing.

Metric	Count
Post-processing Users	250
Post-pre-processing Publications	9.917
Post-processing Terms	72.088
Post-Processing Single Terms	14.744

TABLE 4. Information associated with the Twitter dataset post pre-processing.

Metric	Count
Post-processing Users	59
Post-pre-processing Publications	4.681
Post-processing Terms	25.070
Post-Processing Single Terms	9.914

B. DATA PROCESSING AND ANALYSIS

With the textual data obtained from publications on the social networks Facebook and Twitter, the preprocessing, analysis and the extraction of attributes for the subsequent training and testing of Machine Learning models was performed. This procedure includes in the analysis phase of the textual data the evaluation of the similarity of words between the different languages, using the Word Embeddings model itself. Next, each of the steps of this phase will be discussed.

1) TEXT PREPROCESSING

Textual data from the English language Facebook and Brazilian Portuguese language Twitter datasets were processed using the Python programming language Natural Language Toolkit (NLTK) library [50]. In this sense, the features relative to each language were used to perform stopwords removal, digit and symbol removal, adaptation of common social network abbreviations, uppercase to lowercase transformation, urls removal, other users' mentions and hashtags. Table 3 shows the information from the myPersonality dataset after preprocessing the data.

Regarding the volunteers' data, the information used in the research was only the textual data published, without references to the users' names, mentions or hashtags made, besides the FFM personality patterns associated with the user through the answers to the questionnaire. Thus, a filtering of the Twitter dataset was performed in order to eliminate private profiles, non-existent or no tweets, and to take into account the permanence only of publications with textual content in Brazilian Portuguese language. The details about the terms produced by the participants and also Twitter users, available in the dataset, are shown in Table 4.

2) TEXTUAL DATA ANALYSIS

As Word Embeddings are able to maintain the semantic and syntactic relations of the words in respect to specific contexts, these techniques were used to analyse textual data. Furthermore, the *FastText* algorithm [51] of Word Embeddings is able to deal with words outside the vocabulary, taking into

account the division of words into character units or subwords, which may or may not have semantics.

With the textual data treated, two Word Embeddings models are chosen, each pre-trained with information in a single language, i.e. English or Brazilian Portuguese. In this regard, pre-trained models were chosen using the FastText [51] algorithm with data from the Common Crawl[6] and Wikipedia[7] platforms, having the Continuous Bag of Words (CBOW) settings, 300 dimensions, as well as using the character model *n-gram* of length 5 and context window equal to 5.

After that, the Word Embeddings models were employed to perform analyses on the word coverage of the models with respect to the English and Brazilian Portuguese language datasets. This procedure allowed the evaluation of the mapping of the terms present in the datasets coming from Facebook and Twitter, respectively.

Next, the most frequent terms in the textual content of the Brazilian Portuguese dataset are gathered, and the most similar words to the most frequent ones present in the Brazilian Portuguese Word Embeddings model are identified, in order to identify whether the translations of these terms in the English Word Embeddings model have similarity in the English language, as occurs for the Brazilian Portuguese language. The metric used to measure the similarity between the terms was the *Cosine Similarity*, shown in Equation (1), which is often provided by libraries that provide Word Embeddings algorithms in the Python programming language.

$$S_c = \cos\theta = \frac{A \cdot B}{||A|| \cdot ||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sum_{i=1}^{n} A_i \cdot \sum_{i=1}^{n} B_i} \quad (1)$$

Equation (1) aims to obtain the cosine of an angle θ formed between two distinct word vectors, consisting of the inner product between the vectors and their division by the product of the norm of the vectors. The result coming from the metric is a continuous value in the interval $[-1,1]$, where the closer to 1, the greater the similarity between the terms, while the closer to -1, the greater the dissimilarity between them.

The analysis of the textual data on the most frequent and most similar words can provide indications that there is a relationship between the words pertinent to the different languages through the chosen Word Embeddings model, which suggests the feasibility in the process of generalization of the models considering textual data in different languages between the training and testing stages.

3) FEATURE EXTRACTION

After the analysis of the most frequent words, the feature extraction process was performed using the FastText pre-trained models for the textual data on different languages. This model was also used in [41] to convert Twitter and Facebook English textual data into numerical data, prior to the model training process.

[6]https://commoncrawl.org/
[7]https://pt.wikipedia.org/

The FastText pre-trained models were used to the formation of word vectors based on each of the words of Facebook and Twitter publications, available in the datasets obtained in English and Brazilian Portuguese, respectively. The final word vectors of each publication are formed from the *Word Vector Addition* technique, as verified in [52], which deals with adding all the vectors of each individual word and dividing by the number of words present in the publication.

The word vector data is stored as attributes in additional datasets, which also join the personality values associated with each user across all their publications. The intent of this process is to transform all the terms present in each publication to the users' personality patterns, allowing the subsequent use of Machine Learning algorithms for training models with the Word Embeddings data and predicting personalities based on textual data.

C. MACHINE LEARNING

For training the models, the Sklearn [53] library in Python, which provides Machine Learning algorithms, was used. Thus, the data of word vectors associated with the FFM personality patterns of the users present in the English language Facebook dataset are used. The Machine Learning models used in this work were LR, SGD, AB, SVR and MLP.

Training is performed for five different instances of each model, one for each FFM personality trait. For each user text, the five instances of a chosen model are used, performing predictions in order to obtain the scores for each of the FFM personality traits related to the user.

In addition, hyperparameter search is performed for the evaluation of the best performing configurations using the Grid Search algorithm, 5-fold cross-validation and the Root Mean Squared Error (RMSE) metric. During the evaluations, the RMSE results were converted to a Normalized RMSE (NRMSE), according to the FFM [1], [5] traits scale. The aim was provide an evaluation of the metric in a [0,1] error scale.

Thus, the best performing validation model identified, on average, is trained using all word vectors associated to the English myPersonality dataset texts, considering each of the configurations found through the Grid Search algorithm. Furthermore, the models trained for each of the personality traits are conducted to the Test Stage.

Next, in each subsection, a brief overview will be given about each of the models used, including the hyperparameters of the Sklearn library models considered in this work for the search and optimization algorithm.

1) LINEAR REGRESSION

The LR model seeks to develop a representation of the output value based on a linear combination of input attributes [54]. The intent of the model is to be able to minimize the sum of the residual squares between the actual values of the dataset and the values predicted by the model using a regularization. This model was used in Howlader et al. (2018), which did a comparative study of FFM personality trait predictions

via topics identified in Facebook posts using TF-IDF and extracted with LIWC.

Among all the hyperparameters made available by the LR model in the *Sklearn* library, α was chosen for the phase of searching for the best hyperparameters. This is because it directly influences the regularization present in the process of minimizing the weights, proposing control on the intensity of the reductions performed on the weights assigned to the model inputs during the construction of the model.

2) SUPPORT VECTOR REGRESSOR

SVR is a Machine Learning model responsible for performing a separation of data arranged in space through a hyperplane. To perform this separation, the positions of the support vectors, which are the points in the dataset closest to this hyperplane, are used as parameters [56]. To evaluate the generalization capabilities of the models across different social network posts in automatic personality recognition, Carducci et al. (2018) used SVR to compare its performance with other traditional models for the same task.

Regarding the hyperparameters provided by the *Sklearn* library for the SVR model, the presence of the kernel type is verified, whose function is used to evaluate the data in feature space that allows better partitioning of the sample space using a hyperplane. This is done by applying the kernel function, which provides the transformation of data from low dimensional spaces to a higher dimensional space, turning a nonlinearly separable problem into a linearly separable problem.

In this work, the following types of kernels made available by the library were evaluated (i.e. linear, RBF, sigmoid, polynomial); however, the linear kernel was the one that presented the best results, and therefore, only the hyperparameters of the linear kernel were considered during the evaluations of the results. Another relevant hyperparameter is C, which influences the separation hyperplane of the data samples, allowing the search for an ideal relationship in the definition of the largest minimum margin, capable of adequately separating also the largest possible number of samples by the hyperplane.

3) ADABOOST

AB is a strong model whose representation involves a set composed of weaker Machine Learning models [57]. The idea is to sequentially create the weaker models, so that the result obtained for each one interferes with the following predictions. In [58], data from Facebook social network profiles was used to adjust classification models involving FFM personality traits, as well as seeking to validate the hypothesis that users of similar personality exhibit behavioral patterns with reciprocity characteristics when cooperating via social networks. In this work, the AB built with decision trees was able to outperform the other models evaluated.

The AB regressor model of *Sklearn* has the learning rate hyperparameter, which indicates the weight associated with the influence of each weak estimator on the final decision of the strong estimator. In addition, another relevant hyperparameter available is the number of estimators, which refers to the maximum number of weak estimators to complete the training process.

There is a relationship between the learning rate and the number of estimators, since for low learning rates and therefore little contribution per weak regressor, a number of regressors can be used that makes the smaller contributions more effective for the strong regressor built. The instance also provides the three loss functions presented in the AB model detail. It is emphasized that the weak estimators used in this work were DT regressors, which are the default of the Sklearn library for AB. In addition, the DT estimators use the Mean Squared Error (MSE) as a loss function.

4) STOCHASTIC GRADIENT DESCENT

The SGD used in this work is a model based on Linear Regression, with the differential that it has the characteristics of the optimization algorithm. In this sense, the algorithm seeks to perform a linear combination of the input attributes, besides the minimization of the training regularization error, responsible for measuring the model fit through penalties applied according to the complexity of the model [59]. In Arroju et al. (2015), the SGD was used to perform age, gender, and FFM personality traits recognition in a multilingual setting, already considering two or more languages during model training.

For hyperparameter optimization, the SGD regressor of the Sklearn library allows configuring four different routines[8] for the learning rate, which are named *constant*, *optimal*, *invscaling* and *adaptive*. Other hyperparameters used establish relationships with each of these routines, being: η_0, the initial value of the learning rate for *constant*, *invscaling* and *adaptive*; and α, the constant that influences the regularization strength for *optimal*.

5) MULTILAYER PERCEPTRON

MLP is a neural network architecture in which, from the input data, neurons present in the hidden layers are used to perform nonlinear operations (nonlinear activation functions) and produce a response in the output layer [60]. The MLP neural network was also used in [61], which evaluated the performance of traditional and Deep Learning models for automatic personality recognition in the classification process.

The MLP regressor present in the library *Sklearn* allows the optimization of hyperparameters through changes in the settings that involve the architecture of the network, such as the number of hidden layers, the number of neurons in each layer and the activation function. In addition, other hyperparameters are available to assist in the settings of the backpropagation process, such as the optimizer for updating the neuron weights, the regularization term and the routines

[8] https://scikit-learn.org/stable/modules/sgd.html#id5

constant, *invscaling* and *adaptive* for the learning rate, as well as the initial value of the learning rate.

D. MODEL EVALUATION

Following the model training and validation stages, the Testing Stage occurs, in which the best model trained with English language texts is selected. In order to evaluate the generalization capability of the best model, it is applied to make predictions in Brazilian Portuguese language texts during the Testing Stage. These predictions are carried forward to the result evaluation development stages. The details of the testing and results evaluation steps are shown in the following.

1) TESTING STAGE

After the English results validation is performed during the Training Stage, the Test Stage is performed to do the FFM traits predictions. The aim of this stage is to use the best English trained models for each FFM traits to do predictions based on the FastText Embeddings features extracted from the Brazilian Portuguese Twitter dataset. The results obtained can be used to evaluate these models predictions regarding the real users' personality patterns.

2) EVALUATION of RESULTS

After the models have made predictions through textual data in a language other than the one in which they were trained, that is, Brazilian Portuguese, their performance is evaluated. The objective of this evaluation is to analyze the generalization capability of the models on textual data in different languages between the training and testing stages. The evaluation, in turn, will allow us to verify whether the results are satisfactory when compared to other works that propose the automatic recognition of personalities. To do this, an adaptation of the RMSE metric, the NRMSE metric, is applied to the predictions made by the trained models, as shown in Equation 2.

$$NRMSE = \frac{1}{N} \cdot \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{2}$$

Equation 2 shows the actual value y_i for the personality trait obtained by the form and the inferred value $\hat{y}_i$ by the model, in addition to the total number of samples n from the test dataset. The value N represents the length of the range of variation of the scale defined for the FFM personality traits available in the used datasets. The smaller the error presented by the metric, the closer the inferred values become to the real values and the more adequate is the prediction. This adaptation of the RMSE metric is useful when it is necessary to convert the error values obtained to a predefined scale from 0 to 1. For the datasets used in this work, the values for each FFM personality trait vary on a scale of 1 to 5. Since the scale of variation is known, the length of the interval N can be set equal to 4.

Furthermore, NRMSE allows understanding and unifying the obtained error values involving any interval for each personality trait, including from related works. Due to this, the metric guarantees the conversion of the results in error obtained by related works to a scale from 0 to 1, ensuring the subsequent comparison and discussion of the results. In other works involving the use of regression models for automatic personality recognition, such as [41] and [62], the evaluation metrics RMSE and MSE, respectively, were used to evaluate the performance of the models, which makes it simple to convert both metrics to the NRMSE metric, allowing the analyze and discussion of the results.

IV. RESULTS

Initially, the word similarity analysis was performed in order to verify whether the pre-trained FastText model in English shows similarity between the most frequent terms and their most similar terms, in the same way that occurs in the pre-trained FastText model for Brazilian Portuguese. The terms extracted from the dataset associated with the Brazilian Portuguese language are presented in Table 5, and their translations and percentage of occurrence of similar terms in the English language are shown in Table 6.

Thus, it was noticed on average that about 46% of the terms similar to the most frequent ones extracted in the Brazilian Portuguese language had their English version, similar to the most frequent ones translated into English. Although the results in word similarity were not ideal, they were considered reasonable for studies involving the generalization capability of the models. Thus, the Machine Learning models were evaluated with the 5-fold validation technique and the Grid Search, with the sets of hyperparameters defined in Table 7, to obtain the best model configurations.

The experimental setup involved the use of a notebook computer with Intel(R) Core(TM) i3-6100U CPU 2.30 GHz, Intel(R) HD Graphics 520 GPU, 4GB DDR4 RAM, 500GB HD, and Python programming language version 3.6. The results were collected during one week of processing for the development of the selected models.

Table 8 presents the NRMSE results of the five best performing instances for each Machine Learning model associated with each FFM personality trait. From the results obtained, it was found that the SGD model performed on average 5.23% better compared to SVR, 1.45% better compared to LR, 0.06% better compared to AB, and 2.9% better compared to MLP.

The best α founded to SGD configuration involved the value 10^{-5} to all FFM traits, and the best η_0 founded was 10^{-2} to A and C traits, 10^{-3} to E and O traits and 1 to N trait. In addition, the best learning routine configuration was *invscaling* to A and N traits, and *constant* to E, C and O traits. Other best hyperparameter configurations found for other models have been included in the Supplementary Materials.

Due to the performance results of SGD being the best on average, even with a small difference, it was chosen for the testing phase and subsequent comparison with related works. The selection of this model only, among the five models used, was to facilitate the evaluation of a selected model's ability to

TABLE 5. Top 10 most frequent terms and their similar terms.

Term	Similar terms
vou	ir, pretender, poder, voltar, vamos, prometer, passar, tentar, estar, acabar
gente	molecada, pensar, rapaziada, imaginar, mulherada, molecadinha, pessoa, olhar, achar, andar
acho	acreditar, achar, crer, pensar, entender, admitir, considerar, imaginar, perceber, confessar
casa	residência, mansão, chácara, vizinha, alugar, família, fazenda, cozinha, quintal, cabana
querer	desejar, precisar, gostar, pretender, conseguir, tentar, importar, preferir, esperar, poder
vida	cotidiano, viver, vivência, terrena, história, felicidade, humana, eterna, sobrevivência, trajetória
deus	divindade, castigador, onipotente, divino, demônio, cultuador, oráculo, adorador, mitológico, anjo
cara	babaca, nerd, idiota, sujeito, olhar, malandro, moleque, rapaz, esperto, vergonha
amigo	companheiro, querido, irmão, colega, camarada, primo, tio, parceiro, namorado, vizinho
amo	adorar, odiar, admirar, apaixonar, apreciar, gostar, valorizar, idolatrar, sentir, preferir

TABLE 6. Occurrence of the English translation of the Top 10 most frequent terms and their similar terms.

Term	Similar terms	Percentage
will	go, **intend**, **able**, back, **let**, promise, pass, **try**, **be**, finish	50%
people	**guys**, kids, think, imagine, woman, kids, person, look, find, walk	10%
think	**believe**, find, belief, think, understand, **admit**, **consider**, **imagine**, perceive, confess	40%
home	**residence**, **mansion**, **farmhouse**, neighbor, rent, **family**, **farm**, kitchen, backyard, hut	50%
want	**desire**, **need**, like, **intend**, achieve, **try**, import, prefer, expect, **able**	50%
life	**everyday**, **live**, experience, **earthly**, history, **happiness**, human, **eternal**, **survival**, trajectory	60%
god	**deity**, punisher, **omnipotent**, **divine**, **demon**, culter, **oracle**, worshipper, **mythological**, **angel**	70%
dude	stupid, nerd, idiot, guy, look, rogue, brat, boy, smart, shame	0%
friend	mate, **dear**, **brother**, colleague, **buddy**, **cousin**, **uncle**, partner, boyfriend,**neighbor**	70%
love	**adore**, **hate**, **admire**, impassion, **appreciate**, like, valorize, **idolize**, feel, **prefer**	60%

fit English textual data and to evaluate its performance in the same task with Brazilian Portuguese textual data.

In this context, the selection criterion involved choosing the model with the closest to optimal performance according to the results obtained in the validation phase. The model that best matched NRMSE performance was selected for the English language textual data. The goal was to evaluate whether the model specialized for English in personality recognition is able to provide satisfactory performance for the task for Brazilian Portuguese language textual data.

It is notable that the difference in performance is minimal between the models used for personality recognition using English language textual data. However, the SGD model showed a better average performance when compared to the other models and was therefore chosen for the testing phase.

V. DISCUSSION

The main discussion of the paper involves the feasibility of training a model with textual data from the English language and using it to predict personality traits in the Brazilian Portuguese language. The experiments performed allowed us to verify the best Machine Learning model assigned to the task, i.e. the SGD, as well as its best hyperparameters in the context of the experiments performed. With the best model for the English textual data defined, it was possible to perform new experiments with the Brazilian Portuguese textual data.

As a way to evaluate the generalization capability for FFM traits predictions in Brazilian Portuguese textual data, using the best SGD model trained with English language textual data, the [41] and [62] works were selected. These works were selected in order to make a comparison of the performance results obtained. It is important to emphasize that the works chosen for comparison do not have a methodology similar to the one proposed. This is because they do not explore the capabilities of a model trained with English texts to present

TABLE 7. Sets of hyperparameters defined per model for the search for best configurations.

Model	Hyperparameter	Value
LR	α	1, 2, 3, 5, 8, 10
SVR	C	0.5, 1, 10, 100
AB	learning rate	0.1, 0.3, 0.5, 0.7
	number of estimators	100, 200, 300
	loss function	linear, quadratic, exponential
SGD	learning routine	constant, optimal, invscaling, adaptive
	α	$10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}$
	η_0	$10^{-10}, 10^{-8}, 10^{-5}, 10^{-3}, 10^{-2}, 1, 5, 10$
MLP	number of neurons	50, 100, 150, 200, 250, 300
	activation function	tanh, relu
	optimizer	sgd, adam
	learning routine	constant, adaptive

TABLE 8. NRMSE result of the 5-fold cross-validation process.

Model	O	C	E	A	N	μ
SVR	0,1870	0,1870	0,2190	0,1768	0,1953	0,1930
LR	0,1480	0,1865	0,2195	0,1775	0,1963	0,1856
AB	0,1508	**0,1850**	0,2165	**0,1743**	0,1938	0,1841
SGD	0,1470	0,1853	**0,2153**	0,1753	**0,1915**	**0,1829**
MLP	**0,1455**	0,1875	0,2230	0,1790	0,1988	0,1868

satisfactory performance for the task when applied to Brazilian Portuguese textual data.

In this sense, experiments were conducted to evaluate the generalization capabilities of the models between textual data from different cultures, aiming to solve the problem of the low amount of textual available resources for the FFM personality traits prediction that exist for Brazilian Portuguese. Due to this, the works selected for comparison did not use textual data in different languages, as was used in this proposal, to train the models with data in English language and test their generalization capabilities with data in Brazilian Portuguese language.

The works used as a comparison performed FFM personality traits prediction approaches using regression models with or without the use of textual data. The purpose of the comparison is not to show that the model found in this work is the best among the comparisons for all personality traits of the FFM, but that the results obtained with the approach performed are satisfactory, besides being close to the results obtained by other works related to the task. Because the approach is novel, there are no works that perform a similar procedure. Due to this, there are some limitations in the comparison.

In summary, the main limitations of comparisons are:

(i) The comparative works do not have a methodology similar to the one proposed in this paper, i.e., they do not aim to solve low-resource language problems for FFM trait prediction;

(ii) Different datasets were used by the comparative papers in the model testing stage. Furthermore, no other public datasets in Brazilian Portuguese were identified for performance comparison.

TABLE 9. Comparison of NRMSE results for FFM with related work.

Model	O	C	E	A	N	μ
[41]	0,1293	**0,0908**	**0,1398**	0,1543	**0,1370**	**0,1302**
[62]	0,1725	0,1900	0,2200	0,1975	0,2125	0,1985
SGD	**0,1250**	0,1223	0,1678	**0,0935**	0,2420	0,1501

However, as mentioned above, these chosen works performed FFM traits recognition using Machine Learning regression models. In addition, the results obtained in these works were presented in RMSE and MSE by the authors. These metrics results could be normalized considering the myPersonality dataset scale, which was also used by these works. Table 9 displays the NRMSE results for the FFM of PM in comparison with the works of [41] and [62].

During the experiments performed by [41], a pre-trained FastText model was used as feature extraction method for English textual data obtained from Facebook myPersonality and a Twitter dataset. The extracted features, i.e. English word vectors, were used for training and testing the Machine Learning models. According to the validation experiments performed, SVR was the best performance model. Thus, the comparison of the proposed method performance for SGD model to Brazilian Portuguese in relation to SVR performance to the English-only, which is easier to the models learn similarity relations between words, could shed light about how satisfactory and close are the predictions obtained through texts.

Compared to [41], there were improvements in performance of 39.40% for A and 3.32% for O traits, and an average reduction in performance of 28.61% considering E, N and C traits. In the final average (μ) NRMSE metric for all models, among all personality traits, a decline in performance of 13.26% when compared to [41] was verified. Despite this, the average (μ) results are still satisfactory and close to the results obtained by this work.

The main indication of the better performance obtained by [41] in relation to the SGD model of this proposal is due to the use of only textual data in English, both in the training phase and in the testing phase. The transformations of the words that occur in the English language for the training dataset and, in this same language for the test dataset, lead the model to better understand the formed vectors and their vector meanings. However, this strategy differs from the proposed method, in which the model is fitted for English language textual data and evaluated with Brazilian Portuguese textual data. This is because the models used tend not to recognize exactly the same patterns of the data with which they were

fitted, but rather approximate patterns that may have a similarity relationship.

Additionally, in the experiments performed by [62], the data used was obtained from Twitter users' attributes, such as number of followers, followed, Twitter social network listings, in addition to influence scores, such as Klout and TIME. A single model was used in this work, i.e. Decision Trees regressors with linear models in the leaves using M5' Rules algorithm, which is also a traditional Machine Learning method. The comparison between the proposed method performance for SGD model to Brazilian Portuguese in relation to user attribute method was performed to obtain insights in relation to a textual different method of FFM traits recognition, since this approach does not depends on languages to perform predictions.

In relation to [62], with the exception of the N personality trait, which presented a reduction in performance of 12.19%, it was verified an average improvement of 34.89% in performance, considering A, E, C and O traits. In the final average (μ) NRMSE metric for all models, among all personality traits, it is verified that there was a 24.38% improvement in performance when the SGD is compared to [62]. This shows that the results of the model trained in the English language texts and tested with Portuguese language texts were superior to an approach that use only OSN users' attributes, which is language independent.

It is worth noting that the comparative studies did not evaluate the performance of the models considering data in different languages, being possible to verify that the SGD model trained with data in English language and tested with data in Portuguese language was able to satisfactorily perform the recognition of personalities. Thus, it was proved that: (i) Word Embeddings models can be a sufficient alternative to deal with problems of predicting FFM personality traits on low resources languages, such as Brazilian Portuguese; and (ii) It is possible to use traditional Machine Learning models trained on English language texts to predict personalities through Brazilian Portuguese language texts using the proposed method.

Futhermore, it is highlighted that adversities in the training process of regression models in Machine Learning, such as a low quality and variety of data of continuous nature that denote the personality of individuals, can be detrimental for predictions. As verified in [26], the obstacles in obtaining labeled datasets, as well as in identifying the most relevant attributes for each language and in the preprocessing methods applied, preclude an evident improvement in the predictions. This factor also takes into account the extraction of attributes from textual data in different languages and distinct social networks which, being environments capable of providing users with freedom of expression, have specific vocabularies (i.e. terms, slang, abbreviations) in each different language. Also, the quality of training is also affected by the lack of precision in defining the personality of individuals during data collection through questionnaires, and may be susceptible to human error [63].

In relation to the vectorization techniques, the text transformation into word vectors through Word Embeddings models does not consider the order of the terms present in the users' posts. In this sense, mathematical operations are often performed with the vectors formed, such as the application of the Word Vector Addiction technique, performing the calculation of the average of the positions in the n-dimensional space, according to each term present in the publications submitted to the attribute extraction phase with Word Embeddings [52]. As stated in the related works, the BERT Embeddings models are also available to produce word vectors of sentences and not only of words. However, it is a high computational cost strategy.

As verified in [64], the BERT model is one of the most used state-of-the-art models nowadays, being able to present one of the best results in performance for Text Mining tasks. However, there are drawbacks in its use due to its complexity in training and testing [65]. According to [66], training BERT models on small datasets is not as feasible as a simple model, such as LSTM. BERT consumes large-scale resources, as well as requiring more time during the training stages, especially when a hyperparameter search is considered.

In contrast, traditional Machine Learning models combined with Word Embeddings techniques consume fewer computational resources, require less time to train and are also capable of providing results as satisfactory as BERT. For example, in [67] a study was conducted using the model for refined event classification. It was indicated in the paper that although BERT showed the best results in performance, a traditional Machine Learning SVM model coupled with the traditional TF-IDF technique of Natural Language Processing was able to show results close to BERT. These results are even more attractive, especially when training and evaluation time are taken into consideration when comparing the different models.

Furthermore, the particularities of the terms, present in the Brazilian Portuguese language in OSNs texts, were not present in the training of BERT models available in platforms such as HuggingFace.[9] Due to this, it is not feasible to perform fine-tuning of BERT, as a high computational cost would be demanded to obtain results that might be unsatisfactory in predicting FFM traits. Furthermore, multilingual BERT models may also show results that are not satisfactory when compared to language-specific models. According to [68], multilingual models can still be considered mere substitutes for language-specific models. This is because these models may present inadequate relationships between texts and outputs for solving problems that do not involve many languages.

Due to the aforementioned reasons, BERT models were not used during the experiments in this paper. The intention of the proposed method was to provide a feasible, easily replicable, and computationally simple solution for FFM personality

[9]https://huggingface.co/docs/transformers/model_doc/bert

traits recognition in low resource languages. However, the intention is to use it in future work and compare it with the results obtained in this proposal.

VI. CONCLUSION

In this work, a similarity analysis and a Machine Learning generalization evaluation was performed. The main goal was to answer the main research question: ''Can models trained to predict FFM personality traits using English textual data show satisfactory results when applied to textual data in other languages?''. First, the experiments and comparisons performed demonstrate that it is feasible to use Word Embeddings techniques in order to solve FFM personality traits prediction problems on low resources languages. Futhermore, the proposed method can also present satisfactory and close performance to approaches which has used a single language to the task, in addition to also outperforms approaches which do not rely on language extracted features. Still, the proposed method is innovative and methods similar to the proposal were not found.

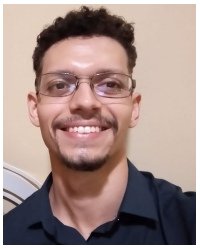
As limitations, it was not found works that proposes similar methods of generalization capabilities evaluation to the task, aiming to deal with low resources languages problems, in order to compare the different approaches. Due to this, a comparison of the proposed method with well-established works in the field was performed, considering two different feature approaches for training the models. The intention was to show that the performance achieved are close to a single-language approach and even better than approaches that do not use language resources for the task. In addition, currently, due to the focus of the work being a viable, easily replicable and computationally simple proposal, more recent models, such as BERT, were not used.

As future work, the goal is to evaluate FFM personality traits in the context of sentence formation using newer Natural Language Processing and Machine Learning models, such as BERT. These models will be used to produce sentence vectors and also will be evaluated in Transfer Learning for FFM traits recognition. Other characteristics will be further explored between English and Brazilian Portuguese language data, addressing a better explanation about the existing relationships between words and sentences of both languages and how they are related to the task. Understanding these relationships could shed light about the relations between high-resource domain languages and low-resource domain languages, such as Brazilian Portuguese. Finally, the intention is to proposes frameworks to deal with low-resource languages problem for FFM personality traits recognition.

REFERENCES

[1] P. Remann and A. Nordin, ''Personality tests in recruitment,'' ResearchGate, Tech. Rep., 2021. [Online]. Available: https://www.researchgate.net/publication/349466573_Personality_tests_in_recruitment
[2] A. Furnham, ''Personality and activity preference,'' *Brit. J. Social Psychol.*, vol. 20, no. 1, pp. 57–68, Feb. 1981.
[3] P. S. Dandannavar, S. R. Mangalwede, and P. M. Kulkarni, ''Social media text—A source for personality prediction,'' in *Proc. Int. Conf. Comput. Techn., Electron. Mech. Syst. (CTEMS)*, Dec. 2018, pp. 62–65.
[4] E. Sharma, R. Mahajan, R. Mahajan, and V. Mansotra, ''Automated personality prediction of social media users: A decade review,'' *Turkish J. Comput. Math. Educ. (TURCOMAT)*, vol. 12, no. 14, pp. 5225–5237, 2021.
[5] T. K. H. Chan, C. M. K. Cheung, and Z. W. Y. Lee, ''Cyberbullying on social networking sites: A literature review and future research directions,'' *Inf. Manage.*, vol. 58, no. 2, Mar. 2021, Art. no. 103411.
[6] O. P. John, R. W. Robins, and L. A. Pervin, *Handbook of Personality: Theory and Research*. New York, NY, USA: Guilford Press, 2010.
[7] O. P. John, A. Angleitner, and F. Ostendorf, ''The lexical approach to personality: A historical review of trait taxonomic research,'' *Eur. J. Personality*, vol. 2, no. 3, pp. 171–203, Sep. 1988.
[8] A. Elngar, N. Jain, D. Sharma, H. Negi, A. Trehan, and A. Srivastava, ''A deep learning based analysis of the big five personality traits from handwriting samples using image processing,'' *J. Inf. Technol. Manage.*, vol. 12, 3–35, 2020.
[9] L. R. Goldberg, ''An alternative 'description of personality': The big-five factor structure,'' *J. Personality Social Psychol.*, vol. 59, no. 6, p. 1216, 1990.
[10] J. Anglim and P. O'connor, ''Measurement and research using the big five, HEXACO, and narrow traits: A primer for researchers and practitioners,'' *Austral. J. Psychol.*, vol. 71, no. 1, pp. 16–25, Mar. 2019.
[11] R. R. McCrae and O. P. John, ''An introduction to the five-factor model and its applications,'' *J. Pers.*, vol. 60, no. 2, pp. 175–215, 1992.
[12] E. C. Tupes and R. E. Christal, ''Recurrent personality factors based on trait ratings,'' *J. Personality*, vol. 60, no. 2, pp. 225–251, Jun. 1992.
[13] J. M. Digman, ''Personality structure: Emergence of the five-factor model,'' *Annu. Rev. Psychol.*, vol. 41, no. 1, pp. 417–440, 1990.
[14] L. R. Goldberg, ''The structure of phenotypic personality traits,'' *Amer. Psychologist*, vol. 48, no. 1, p. 26, 1993.
[15] L. R. Goldberg, ''A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models,'' *Personality Psychol. Eur.*, vol. 7, no. 1, pp. 7–28, 1999.
[16] L. R. Goldberg, J. A. Johnson, H. W. Eber, R. Hogan, M. C. Ashton, C. R. Cloninger, and H. G. Gough, ''The international personality item pool and the future of public-domain personality measures,'' *J. Res. Personality*, vol. 40, no. 1, pp. 84–96, Feb. 2006.
[17] L. Qiu, H. Lin, J. Ramsay, and F. Yang, ''You are what you tweet: Personality expression and perception on Twitter,'' *J. Res. Personality*, vol. 46, no. 6, pp. 710–718, 2012.
[18] J. M. Balmaceda, S. Schiaffino, and D. Godoy, ''How do personality traits affect communication among users in online social networks?'' *Online Inf. Rev.*, vol. 38, no. 1, pp. 136–153, Jan. 2014.
[19] H. Tankovska, ''Social media—Statistics & facts,'' Statista, Tech. Rep., 2021. [Online]. Available: https://www.statista.com/statistics/1106343/social-usage-increase-due-to-coronavirus-home-usa/
[20] S. Dixon, ''Most popular social networks worldwide as of January 2022, ranked by number of monthly active users,'' Statista, Tech. Rep., 2022. [Online]. Available: https://www.statista.com/statistics/454772/number-social-media-user-worldwide-region/?locale=en
[21] R. Festl and T. Quandt, ''Social relations and cyberbullying: The influence of individual and structural attributes on victimization and perpetration via the internet,'' *Hum. Commun. Res.*, vol. 39, no. 1, pp. 101–126, Jan. 2013.
[22] C. M. Kokkinos, N. Antoniadou, and A. Markos, ''Cyber-bullying: An investigation of the psychological profile of university student participants,'' *J. Appl. Develop. Psychol.*, vol. 35, no. 3, pp. 204–214, May 2014.
[23] S. You and S. A. Lim, ''Longitudinal predictors of cyberbullying perpetration: Evidence from Korean middle school students,'' *Personality Individual Differences*, vol. 89, pp. 172–176, Jan. 2016.
[24] S. D. Gosling, A. A. Augustine, S. Vazire, N. Holtzman, and S. Gaddis, ''Manifestations of personality in online social networks: Self-reported Facebook-related behaviors and observable profile information,'' *Cyberpsychol., Behav., Social Netw.*, vol. 14, no. 9, pp. 483–488, Sep. 2011.
[25] R. Buettner, ''Predicting user behavior in electronic markets based on personality-mining in large online social networks,'' *Electron. Markets*, vol. 27, no. 3, pp. 247–265, Aug. 2017.
[26] V. Ong, A. D. S. Rahmanto, Williem, and D. Suhartono, ''Exploring personality prediction from text on social media: A literature review,'' *Internetworking Indonesia*, vol. 9, no. 1, pp. 65–70, 2017.
[27] J. W. Pennebaker, M. E. Francis, and R. J. Booth, ''Linguistic inquiry and word count: LIWC 2001,'' *Mahway, Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001. 2001.
[28] W. R. Dos Santos, R. M. S. Ramos, and I. Paraboni, ''Computational personality recognition from Facebook text: Psycholinguistic features, words and facets,'' *New Rev. Hypermedia Multimedia*, vol. 25, no. 4, pp. 268–287, Oct. 2019.

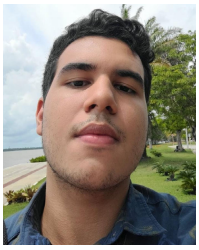
[29] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *J. Artif. Intell. Res.*, vol. 30, pp. 457–500, Nov. 2007.
[30] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from Twitter," in *Proc. IEEE 3rd Int. Conf. Privacy, Secur., Risk Trust IEEE 3rd Int. Conf. Social Comput.*, Oct. 2011, pp. 149–156.
[31] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PLoS ONE*, vol. 8, no. 9, Sep. 2013, Art. no. e73791.
[32] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Personality predictions based on user behavior on the Facebook social media platform," *IEEE Access*, vol. 6, pp. 61959–61969, 2018.
[33] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
[34] F. B. Siddique and P. Fung, "Bilingual word embeddings for cross-lingual personality recognition using convolutional neural nets," *Learn.*, vol. 21, p. 22, Aug. 2017.
[35] F. Liu, J. Perez, and S. Nowson, "A recurrent and compositional model for personality trait recognition from short texts," in *Proc. Workshop Comput. Model. People's Opinions, Personality, Emotions Social Media (PEOPLES)*, 2016, pp. 20–29.
[36] X. Sun, B. Liu, Q. Meng, J. Cao, J. Luo, and H. Yin, "Group-level personality detection based on text generated networks," *World Wide Web*, vol. 23, no. 3, pp. 1887–1906, 2019.
[37] J. Yu and K. Markov, "Deep learning based personality recognition from Facebook status updates," in *Proc. IEEE 8th Int. Conf. Awareness Sci. Technol. (iCAST)*, Nov. 2017, pp. 383–387.
[38] F. O. L. Pabón and J. R. O. Arroyave, "Automatic personality evaluation from transliterations of YouTube vlogs using classical and state of the art word embeddings," *Ingeniería e Investigación*, vol. 42, no. 2, p. 10, 2022.
[39] Z. Ren, Q. Shen, X. Diao, and H. Xu, "A sentiment-aware deep learning approach for personality detection from text," *Inf. Process. Manage.*, vol. 58, no. 3, May 2021, Art. no. 102532.
[40] H. Jiang, X. Zhang, and J. D. Choi, "Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings (student abstract)," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 13821–13822.
[41] G. Carducci, G. Rizzo, D. Monti, E. Palumbo, and M. Morisio, "TwitPersonality: Computing personality traits from tweets using word embeddings and supervised learning," *Information*, vol. 9, no. 5, p. 127, May 2018.
[42] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intell. Syst.*, vol. 32, no. 2, pp. 74–79, Mar. 2017.
[43] X. Sun, B. Liu, J. Cao, J. Luo, and X. Shen, "Who am I? Personality detection based on deep learning for texts," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.
[44] F. B. Siddique, D. Bertero, and P. Fung, "GlobalTrait: Personality alignment of multilingual word embeddings," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 7015–7022.
[45] M. Arroju, A. Hassan, and G. Farnadi, "Age, gender and personality recognition using tweets in a multilingual setting," in *Proc. 6th Conf. Labs Eval. Forum (CLEF), Exp. IR Meets Multilinguality, Multimodality, Interact.*, 2015, pp. 23–31.
[46] W. Maharani and V. Effendy, "Big five personality prediction based in Indonesian tweets using machine learning methods," *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 12, no. 2, p. 1973, Apr. 2022.
[47] F. Celli, F. Pianesi, D. Stillwell, and M. Kosinski, "Workshop on computational personality recognition: Shared task," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 7, 2013, pp. 2–5.
[48] *Free Open-Source Bigfive Personality Traits Test—Big Five*, Rubynor, 2020. [Online]. Available: https://bigfive-test.com
[49] J. A. Johnson, "Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the IPIP-NEO-120," *J. Res. Personality*, vol. 51, pp. 78–89, Aug. 2014.
[50] S. Bird, E. Klein, and E. Loper, *Natural Language Processing With Python: Analyzing Text With the Natural Language Toolkit*. Sebastopol, CA, USA: O'Reilly Media, 2009.
[51] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Jun. 2016.
[52] I. B. Drexel, "Feature engineering and word embedding impacts for automatic personality detection on instant message," in *Proc. Int. Conf. Inf. Manage. Technol. (ICIMTech)*, Aug. 2019, pp. 155–159.
[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 10, pp. 2825–2830, Jul. 2017.
[54] A. Y. Ng, "Feature selection, L_1 vs. L_2 regularization, and rotational invariance," in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*, 2004, p. 78.
[55] P. Howlader, K. K. Pal, A. Cuzzocrea, and S. D. M. Kumar, "Predicting Facebook-users' personality based on status and linguistic features via flexible regression analysis techniques," in *Proc. 33rd Annu. ACM Symp. Appl. Comput.*, Apr. 2018, pp. 339–345.
[56] V. Vapnik, S. Golowich, and A. Smola, "Support vector method for function approximation, regression estimation, and signal processing," in *Proc. Adv. Neural Inf. Process. Syst.*, 1997, pp. 281–287.
[57] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, pp. 119–139, Aug. 1995.
[58] A. Souri, S. Hosseinpour, and A. M. Rahmani, "Personality classification based on profiles of social networks' users and the five-factor model of personality," *Hum.-Centric Comput. Inf. Sci.*, vol. 8, no. 1, pp. 1–15, Dec. 2018.
[59] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*, 2004, p. 116.
[60] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, Oct. 1986.
[61] H. T. Tandera, D. Suhartono, R. Wongso, and Y. L. Prasetio, "Personality prediction system from Facebook users," *Proc. Comput. Sci.*, vol. 116, pp. 604–611, Jan. 2017.
[62] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, "Our Twitter profiles, our selves: Predicting personality with Twitter," in *Proc. IEEE 3rd Int. Conf. Privacy, Secur., Risk Trust IEEE 3rd Int. Conf. Social Comput.*, Oct. 2011, pp. 180–185.
[63] A. K. Tripathi, S. Hossain, V. Singh, and P. Atrey, "Personality prediction with social behavior by analyzing social media data—A survey," Dept. Appl. Comput. Sci., Univ. Winnipeg, Tech. Rep., 2010. [Online]. Available: http://www.cs.albany.edu/~patrey/ICSI660-445/project/Survey_sample_report.pdf
[64] S. González-Carvajal and E. C. Garrido-Merchán, "Comparing BERT against traditional machine learning text classification," 2020, *arXiv:2005.13012*.
[65] A. Ettinger, "What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 34–48, Dec. 2020.
[66] A. Ezen-Can, "A comparison of LSTM and BERT for small corpus," 2020, *arXiv:2009.05451*.
[67] J. Piskorski, J. Haneczok, and G. Jacquet, "New benchmark corpus and models for fine-grained event classification: To BERT or not to BERT? in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 6663–6678.
[68] S. Rönnqvist, J. Kanerva, T. Salakoski, and F. Ginter, "Is multilingual BERT fluent in language generation?" 2019, *arXiv:1910.03806*.



ADRIANO MADUREIRA DOS SANTOS received the degree in computer engineering from the Federal University of Pará (UFPA), in 2022. He is currently pursuing the master's degree in applied computing with the Graduate Program in Electrical Engineering through the Institute of Technology (ITEC) of UFPA, where he is working in computer vision and data mining projects. He contributed to scientific research in the Institutional Program of Scientific Initiation Scholarships (PIBIC), where he was working in social network analysis and text mining projects. He is a CNPq Technological Development and Innovative Extension Grant Holder of the Sectoral Funds for Human Resources (SET) Modality, dealing with education projects through the use of artificial intelligence. He has publications in the areas of social network analysis, text mining, and gamification.

FLÁVIO RAFAEL TRINDADE MOURA is currently pursuing the bachelor's degree in computer engineering with the Federal University of Pará (UFPA). He is a member of the Operational Research Laboratory (LPO) and contributed to scientific research as a fellow of the Institutional Program of Scholarships for Extension (PIBEX-UFPA), where he is involved in research related to artificial intelligence, computer vision, teaching, and machine learning. He is also a Grant Holder of the Technological Development and Innovative Extension (modality fixation and training of human resources sector funds (SET)) and a Scholarship Holder of the Project of Territorial Innovation in the Remnant Community of Quilombo de Igarapé Preto, Oeiras do Pará (PA), for young people and adults.

LYANH VINICIOS LOPES PINTO is currently pursuing the degree in computer engineering with the Institute of Technology, Federal University of Pará (UFPA). He was a scholarship holder in extension projects in the area of artificial intelligence and its subareas. He is currently a Grantee with Pró-reitoria de Extensão (PIBEX), in the 2-D unit game development area.



ANDRÉ VINICIUS NEVES ALVES is currently pursuing the bachelor's degree in telecommunications engineering with the Federal University of Pará (UFPA). He is a member and a Researcher with the Operational Research Laboratory (LPO-UFPA), where he is involved in the areas of web applications and artificial intelligence. He is a fellow of the EduTech Amazon Project: use of artificial intelligence for geometric pattern recognition in streaming in an application for teaching basic education through the App Geometricando 2.0 VR and AR. He received a grant from the CNPq/SEMPI/MCTI edict no. 021/2021-RHAE Program-Line 1-Innovative Companies of CNPq.

KARLA FIGUEIREDO received the B.Sc. degree in electrical engineering from the Federal University of Rio de Janeiro (UFRJ), Brazil, and the M.Sc. and Ph.D. degrees in computer science from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio). She is currently the Head of the Computational Intelligence and Robotics Laboratory (LIRA) developing more than 30 projects in computational intelligence. She has supervised more than 20 Ph.D. theses and M.Sc. dissertations. She is the author of ten book chapters and more than 50 scientific articles in the areas of soft computing and machine learning. Her research interests include computational intelligence methods and applications, including neural networks, fuzzy logic, hybrid intelligent systems, robotics, and intelligent agents, applied to decision support systems, pattern classification, time-series forecasting, natural language processing, control, optimization, and data mining.

FERNANDO AUGUSTO RIBEIRO COSTA received the degree in geography from the Federal Institute of Education, Science and Technology of Pará, and the bachelor's degree in tourism from the Federal University of Pará (UFPA). He is currently pursuing the master's degree with the Graduate Program in Sustainable Development of the Wet Tropic, Center for High Amazon Studies (PDTU/NAEA/UFPA). He is a Specialist in scientific communication with Amazon through the International Program for the Training of Specialists in the development of Amazonian areas, Center for High Amazon Studies (FIPAM/NAEA/UFPA). He is also developing research on the consequences of violence among the population of peripheral neighborhoods in the Metropolitan Region, Belém, Pará. He has publications in the areas of scientific communication, culture, artificial intelligence, and violence against women.

MARCOS CÉSAR DA ROCHA SERUFFO received the Technology's degree in data processing from the University Center of the State of Pará (CESUPA), in 2004, and the master's degree in computer science and the Ph.D. degree in electrical engineering with an emphasis in applied computing from the Federal University of Pará (UFPA), in 2008. He is currently an Associate Professor with UFPA. He is also a Professor of anthropic studies with the Amazon Graduate Program (PPGEAA) and the Electrical Engineering Graduate Program (PPGEE). He is a Researcher with the Operational Research Laboratory (LPO). He is the author of more than 80 scientific articles. His research interests include computing science, social technologies, digital TV, computer networks, access technologies, and human–computer interaction.