ISSN 2007-9737

Creation of a Corpus in Spanish for the Recognition of Personality Traits

Víctor Manuel Bátiz-Beltrán, María Lucía Barrón-Estrada*, Ramón Zatarain-Cabada, Jonathan Iván Roldán-Arana

Tecnológico Nacional de México Campus Culiacán, Posgrado e Investigación, Culiacán, Sinaloa, Mexico

{victor.bb, lucia.be, ramon.zc}@culiacan.tecnm.mx, jonathan_roldan@itculiacan.edu.mx

Abstract. Automatic personality recognition is a research area that has become very important in recent years. Currently there is research with different approaches that seek to automatically recognize personality traits by means of text. There are methods that use texts from voice transcriptions or texts written by people to determine whether an individual has a certain personality trait. These methods are based on machine learning and deep learning algorithms. A key element for the construction of such models is to have a data set (corpus) for training and optimization. This paper presents the creation of a corpus called PersonText, which contains 213 texts in Spanish with their respective labels related to the presence or absence of the personality traits of the Big-Five model, as well as the scores obtained by the participants in a standardized personality test. The main motivation for the creation of this corpus was the limited existence of corpora of texts in Spanish focused on personality recognition. The information was obtained from the platform called PersonApp, used for data collection based on standardized personality tests and videos of the participants. Additionally, to evaluate the corpus, tests were performed with different machine learning and deep learning models. The results obtained are promising and validate the relevance of the corpus built to address the task of automatic personality recognition.

Keywords. Big-Five, corpus, personality, machine learning, deep learning, machine recognition, machine learning, machine recognition.

1 Introduction

Personality is a set of traits that determine the way a person acts or makes decisions. These traits come to be aspects such as emotions, feelings and pattern characteristics of attitudes and thoughts [1]. Personality is what induces us to make decisions, such as choosing the type of clothes we wear every day, the items we buy at the supermarket, the hobbies we have, the places we visit or even the social events we attend. Therefore, automatic recognition of personality traits is important because it allows us to solve problems more efficiently, such as placing the right person in the right workplace or improving educational processes for students according to their personality.

Currently, two of the most common and widely used models for personality recognition based on written tests are the Myers-Briggs Type Indicator (MBTI) model, which is a self-reported questionnaire that aims to assess how individuals perceive their surroundings and make decisions. [2] and Big-Five personality traits [3] (also called OCEAN)

The MBTI test is not accepted in scientific circles due to a lack of evidence demonstrating its reliability as a personality test. In the other hand, the Big-Five model is the most popular and accepted model by the scientific community and is the one used as the basis for this research. In this model the five major personality traits are represented by Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

To automatically recognize the personality of an individual, machine learning (ML) or deep learning (DL) methods are used, which in turn require data sets or text corpora for training the classifiers that will automatically determine the presence or absence of each personality trait. There are

ISSN 2007-9737

corpora in languages other than Spanish for personality recognition, as is the case of the myPersonality corpus, which was the result of a project based on a Facebook application that allowed participants to take a personality test and give their consent for the program to record their profile information [4].

Another corpus with data based on personality traits is the b5 corpus [5], which is a collection of controlled and free texts (without specific subject matter) produced in different communicative tasks (e.g., referential, or descriptive), and accompanied by personality inventories of their authors and other demographic data.

When approaching the task of automatic personality recognition based on Spanish texts, one of the problems found is the low number of corpora based on Spanish texts oriented to personality recognition.

In this research work the following research question is addressed:

– Is it possible to build ML and DL models for automatic personality trait recognition that obtain a performance comparable to the state of the art when applied to a Spanish corpus constructed from textual transcripts of Spanish videos and their corresponding personality trait labels based on a standardized personality test?

The main contribution of this work is the creation of a corpus of texts in Spanish language, which is based on the Big-Five model for personality traits using a novel methodology for its construction, as well as the evaluation of the corpus by means of classification models by ML and DL methods.

The data were obtained from PersonApp [6], a platform developed internally by our research group that aims to collect data through videos and standardized personality tests of the participants, based on the questions available in the International Personality Item Pool (IPIP) [7]. In addition, several classification models were developed and optimized to test the corpus.

This paper is structured in the following order: in Section 2 we present related work in text-based personality recognition; Section 3 presents the methodology implemented for this study; Section 4 shows the results with the classification models and their discussion; finally, Section 5 describes the conclusions and future work.

2 Related Work

This section describes related work in the field of personality recognition using text. The section is divided into two parts: the first part presents works on creating a text dataset (corpus) based on personality recognition models. The second part describes works on automatic recognition where datasets are used for training and classification of personality in text.

2.1 Text Corpora About Personality

There are several corpora used for text-based personality prediction. One of the first such corpus is called essays [8], created by Pennebaker and King, which is a dataset containing approximately 2400 stream-of-consciousness essays, labeled with the personality traits of the authors of each piece of writing.

The corpus is organized as follows:

– Author identifier.
– Text of the authors' stream-of-consciousness essays.
– Personality traits labeled in two classes (yes and no, to indicate presence or absence of each of the Big-Five traits). The field titles for the personality traits are EXT (Extraversion), NEU (Neuroticism), AGR (Agreeableness), CON (Conscientiousness) and OPN (Openness to Experience). The questionnaire used to collect personality traits was the Five-Factor Inventory [9].

In the work of Kim & Walter [10], the corpus personae used for the prediction of authorship attribution and author personality is presented. The corpus consists of essays written by 145 authors (approximately 1,400 words each), as well as personality profiles based on the MBTI test.

This test employs four scales, defined as opposing pairs between eight categories: introversion-extraversion, sensing-intuition, thinking-feeling, and judgment-perception.

Each category is symbolized by a letter, and the result is expressed as a combination of four letters,

ISSN 2007-9737

Table 1. Text corpora based on Big Five model

Corpus	Size (number of records)	Language	Labels	Measured values of personality traits
myPersonality subcorpus [13]	10,000	English	Big-Five (y/n)	0 to 5
PAN-AP-2015 [11]	726	English, Spanish, Italian and Dutch	Big-Five (numeric)	-0.5 to +0.5
b5-post [5]	194,382	Portuguese	Big-Five (yes/no)	-
Essays [8]	2,400	English	Big-Five (y/n)	-
HWxPI [14]	836	Spanish	Big-Five (numeric)	0 or 1

among the sixteen possible, which is intended to define the personality of the subject [2].

There are also corpora whose data were collected from the Facebook application. Among these we can mention the b5 corpus which is a corpus in Portuguese language. The b5 corpus is composed of several subcorpora, being the b5-post subcorpus the largest, since it has 194,382 sentences from 1,019 users [5]. The subcorpus was constructed for the purpose of conducting research and developing computational models for the recognition of personality traits and the profiling of authors.

In [11], Rangel et al. mention the PAN-AP-2015 corpus which is composed of texts collected from the social network Twitter (now known as X), from 726 users, in four languages: English, Spanish, Italian and Dutch. The texts are distributed as follows: 336 are in English, 228 are in Spanish, 86 are in Italian and 76 are in Dutch. The assessment of personality traits was conducted via the BFI-10 online test [12], which was completed by the participants themselves. The corpus is annotated with gender and personality traits in normalized values between -0.5 and +0.5.

Similarly, myPersonality is one of the most popular and widely used corpora for text-based personality recognition tests. This corpus is a database that was constructed by collecting information from Facebook, and where the data are labeled similarly to the essays dataset, with binary values (y/n) to indicate whether the trait is present or not, as well as with their respective personality measurement values obtained by IPIP tests [7] corresponding to each author. In order to assess the performance of the automatic recognition models presented in this research, a subset of the myPersonality corpus, as proposed in [13] by Celli et al., was utilized.

In the Spanish language in [14], the authors present the HWxPI corpus consisting of handwritten essays for personality identification. The corpus contains information from 836 participants. Two modalities of each handwritten text are provided: the manually transcribed essay and the scanned image of the essay.

During the data collection phase, the researchers employed a psychological instrument, the TIPI (Ten-Item Personality Inventory) [15], to ascertain each subject's personality characteristics according to the Big Five Model. The values for each personality trait are indicated in binary form. The presence of the trait is denoted by 1 and the absence by 0. A summary of the content of the corpora related to the Big-Five model for personality recognition is shown in Table 1.

2.2 Automatic Personality Recognition by Text

In recent years, several research works have been carried out for automatic personality recognition from text classification, in which ML or DL techniques are used, taking some text corpus for training these methods. For this purpose, different approaches are used, such as using texts from voice transcripts or texts written by the person or published in social networks, from which it can be determined whether a person has a certain personality trait [16, 17, 18].

Majumder's work [19] shows a method where, starting from the essay's corpus (used as training data), they use a binary classification convolutional neural network (CNN) for each personality trait with the same structure, whose purpose is to classify whether the personality trait is present or not.

The process starts with a preprocessing and filtering of the input data, where all sentences of each trial are transformed into n-gram feature vectors, followed by a feature extraction, where the vectors are concatenated with the per-word semantic features and with the François Mairesse features [20], which is a set of document-level features for personality detection.

The result is a variable length representation, which is fed into the CNN where the classification is performed. The accuracy results obtained for the personality traits are in the range of 56.71% to 62.68%.

Tandera et al. [21] conducted research based on the Big-Five personality model. This research uses the myPersonality corpus. In this work they tested traditional ML and DL methods. In ML models they used algorithms such as Support Vector Machine (SVM), Logistic Regression, Naive Bayes, Gradient Boosting and Linear Discriminant Analysis (LDA) and as DL models they tested MLP (multilayer perceptron), GRU (Gated Recurrent Unit), LSTM (long Short-Term Memory), CNN 1D (one-dimensional convolutional neural network), and a combination of the latter two (LSTM + CNN 1D).

The authors conclude that their experiments results demonstrate that deep learning techniques can enhance the accuracy of recognition models. However, they acknowledge that for certain personality traits, the obtained accuracy values were relatively low. The evaluated models achieved an accuracy in the range of 68.63% to 74.17%.

In the work of Liu et al. [22], short text personality is determined using deep learning models from the PAN 2015 Author Profiling task dataset [11], which is an English language database gathered from Twitter, containing approximately 14,000 tweets and the five personality traits labeled and evaluated each with values between -0.5 and 0.5 corresponding to 152 users.

The evaluation metric used was Root Mean Square Error (RMSE), obtaining results in the range of 0.109 to 0.167 when analyzing tweets at the user level and in the range of 0.127 to 0.189 in the analysis of individual tweets. The authors conclude that their methodology achieves results comparable to those of the state of the art in predicting personality traits.

3 Description of the Methodology

This section shows the methodology for the construction of the PersonText corpus. It is worth mentioning that the texts of the participants' opinions, as well as the content of the corpus is completely anonymous, respecting the privacy of the participants. To guarantee this, each text was reviewed and, in the parts where a name was mentioned, it was omitted or replaced by the following word: "$NAME$". Additionally, the architecture of the ML and DL models used to test the corpus is described.

Participants register and log on to the PersonApp platform, where they take the standardized personality test and record videos talking about a topic of their choice. The audio is then extracted from the videos and converted to text. On the other hand, the results of the standardized test are evaluated, and labels of presence and absence of each trait are assigned based on the values obtained in the test.

Finally, the texts, the generic information of the participant, the values obtained in the test and the presence and absence labels of each trait are merged to generate the corpus. Figure 1 shows the general diagram of the process used to build the corpus.

3.1 PersonApp

PersonApp is a multiplatform system developed internally by the research group of the Instituto Tecnológico de Culiacán that aims to collect data such as videos and IPIP standardized personality test results from participants [6].

This platform was created to evaluate the personality of each participant through a standardized personality test made up of 50 IPIP items and to obtain values for each of the five personality traits of the Big-Five model.

In addition, the platform enables participants to record videos in which they are invited to discuss a topic of their choosing. They are encouraged to

ISSN 2007-9737

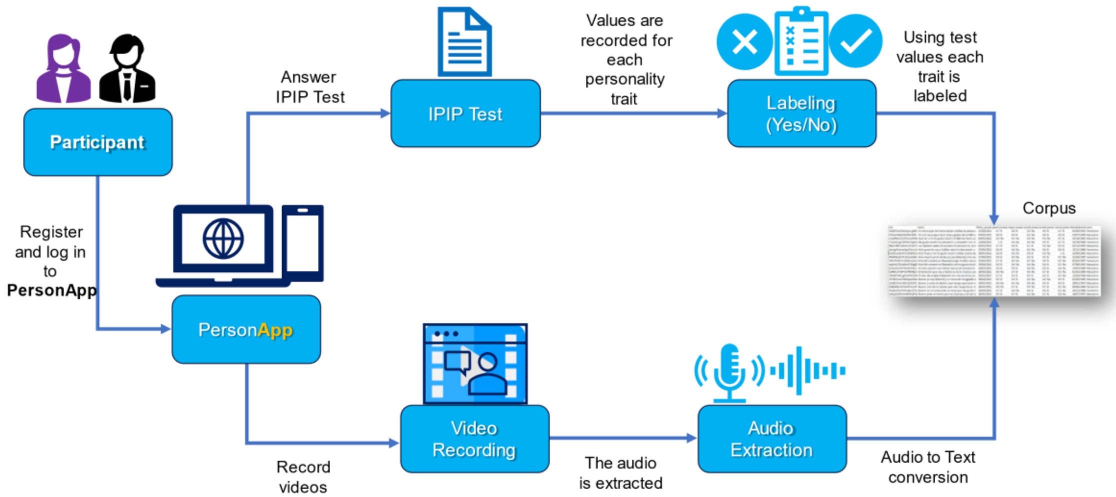

Fig. 1. General diagram of the construction of the PersonText corpus

imagine that they are presenting this topic to their friends in a Zoom or Teams meeting.

Another scenario is to imagine themselves in an online interview and discuss their strengths and weaknesses. Additionally, they are prompted to consider presenting a topic on a streaming platform such as YouTube, Facebook, or Twitch.

3.2 Compilation of Texts from the Videos

For the generation of the texts, the videos of the participants were taken from the PersonApp platform. In these videos the participants express themselves freely on various topics. About 700 videos were processed.

The extraction of the texts in the videos was carried out automatically using the Speech Recognition library [23].

Once the texts were obtained, the research team reviewed them to validate the audio-to-text conversion and correct details that occurred during the process.

In addition to the videos, the platform has all the participants' data such as name, gender, email, age, and the results of the IPIP tests consisting of the scores obtained on a scale between 0 and 1 corresponding to each of the five personality factors.

3.3 Structure of the PersonText Corpus

PersonText is a Spanish-language corpus containing texts of comments from different individuals, as well as their corresponding personality trait values. The inventory with the personality trait values accompanying the texts is the result of the 50-item IPIP test taken by each participant. The PersonText corpus was created in a CSV (comma-separated values) file and has 213 texts in Spanish.

The corpus data are stacked and divided into sections ranging from the user, the text of the responses of each experiment, followed by the IPIP test result values, as well as the presence label (Yes/No) of the personality trait. Table 2 shows an excerpt from the PersonText corpus.

The corpus has presence labels for each Personality trait. These were defined based on the IPIP test scores, where if the assessment result was higher than 0.6 it was labeled with a "Sí" (Yes in English), otherwise with a "No".

3.4 Personality Classification Models

This section describes various methods used for automatic personality recognition. It describes the preprocessing performed on the corpus, and the machine learning and deep learning models used to perform the binary (Yes/No) classification.

ISSN 2007-9737

Table 2. Excerpt from the PersonText corpus

UID	Text	TestDate	OpnV	OpnT	ConV	ConT	ExtV	ExtT	AgrV	AgrT	NeuV	NeuT	Sex	Age
n84XTQwKheUqUcjjd0OIusWEAtg1	Un tema que me tiene dando vueltas la cabeza es que ya viene el día del niño y me gustaría como que hacerle un día especial al niño quisiera decorarle una pared quisiera hacer bollitos para ese día comprarle dulces que haya música para que él esté feliz quisiera que hiciéramos en esa semana actividades con el niño como salir al parque hacerle un día un pastel o algo al niño me gustaría que hiciéramos una video llamada que puedan estar	23/04/2021	0.7	Sí	0.8	Sí	0.6	No	0.9	Sí	0.7	No	Femenino	29
iXWwNMyEN5OBn4F6nMQnoqxNoki2	Si a mí me preguntaran ¿Qué jugador de la NBA es mi favorito? pues por obvias razones es Stephen Curry es más que nada un botador hay más habilidosos en su posición como James Hardy muy buen botador creo que mucho mejor botador que Curry pero lo que hace Curry es sin esfuerzo siempre lo hace sin esfuerzo es natural mientras que Hardy es mucho más habilidoso	04/03/2021	0.8	Sí	0.8	Sí	0.3	No	0.8	Sí	0.9	No	Masculino	26

3.4.1 Text Preprocessing

The preprocessing performed on each text includes operations such as removing Web addresses and line breaks, converting words to lowercase, omitting numbers, discarding non-alphanumeric characters (such as single quotation marks), removing blank spaces, filtering words that do not add value or meaning to the sentence (Stop Words), performing lemmatization on each word (reducing words to their root or base form), and finally removing punctuation marks.

3.4.2. Architecture of Machine and Deep Learning Models

The Keras library with the Python programming language was used to create the models. As a first step, the texts included in the PersonText corpus described in the previous section were preprocessed. Then, features were extracted from the corpus using the Bag of Words method. This method counts the number of times words from a given set (bag) appear in a text document, without considering grammar or word order.

Once the input data was ready, the deep learning model was configured for which different combinations of hyperparameters (such as the number of layers, the number of input and output neurons, and the type of layer in the activation function) were considered. In the latter, softmax was used since it is recommended for classification issues.

In the loss function, we worked with binary cross entropy because this function is frequently used in binary classification problems and in our research work, we seek to recognize Yes/No classes. In addition, different tests were performed with different types of optimizers, finally choosing the RMSprop optimizer, since this optimizer adapts the learning rate in small batches and with it, we obtained better results.

This research work examines several neural networks, including the multilayer perceptron (MLP). The model architecture consists of an embedding layer as the input layer, followed by a Flatten layer and then a dense layer with a ReLU activation function. Finally, as output layer it presents a dense layer with softmax activation function.

ISSN 2007-9737

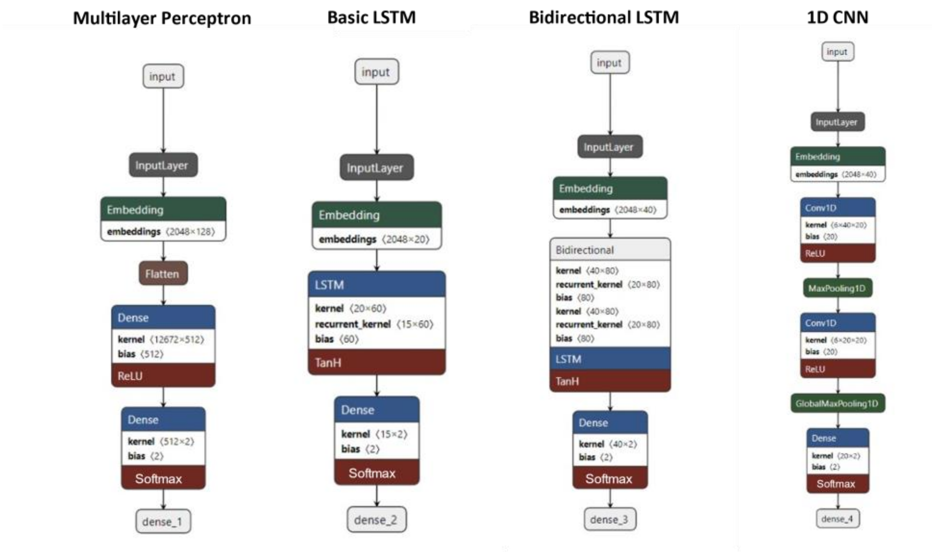

Fig. 2. Graphical view of the configuration of the machine and deep learning models

Training was also done with a basic LSTM neural network which is a sequential model with an input layer. The classifier receives the text in an embedding layer as input layer. In the next LSTM layer, we have 15 neurons and a dropout considered for the output of the embedding layer of 0.5. Finally, we have a dense layer with two output neurons with softmax activation function.

We also worked with a Bidirectional LSTM which is also a sequential model where the layers selected for this network are: embedding layer as input layer, followed by the Bidirectional LSTM layer, and Dense layer with softmax activation function.

Finally, a one-dimensional Convolutional Model (1D CNN) was also configured, which has an embedding layer as input, to which a 1D convolutional layer with ReLU activation is added, followed by a 1D max-pooling layer, a second 1D convolutional layer like the first one, then a global max-pooling layer and finally a Dense layer as output. This layer has two units and uses the softmax activation function to deliver probabilities for two classes.

Figure 2 shows the architectures of the 4 personality classification models used in the PersonText corpus tests.

4 Results and Discussion

Currently the PersonText corpus contains 213 text records with their respective labels and inventory of IPIP test results. Figure 3 shows the distribution of the classes for each personality trait. It is observed that the agreeableness trait is the most unbalanced. On the other hand, the traits of openness and responsibility are the most balanced.

For testing purposes, the PersonText corpus was divided into 75% for training and 25% for validation. As discussed in subsection 3.4.1, before performing the corpus division, preprocessing and cleaning of the corpus was previously performed. For training, 50 epochs were used, where the metric to be monitored was accuracy, keeping the best model found.

4.1 Results

To evaluate the PersonText corpus, different tests were performed with several machine learning and deep learning models. For this purpose, the metrics used in this research were accuracy, which represents the ratio between the number of correct predictions and the total number of input samples, and precision, which consists of determining the

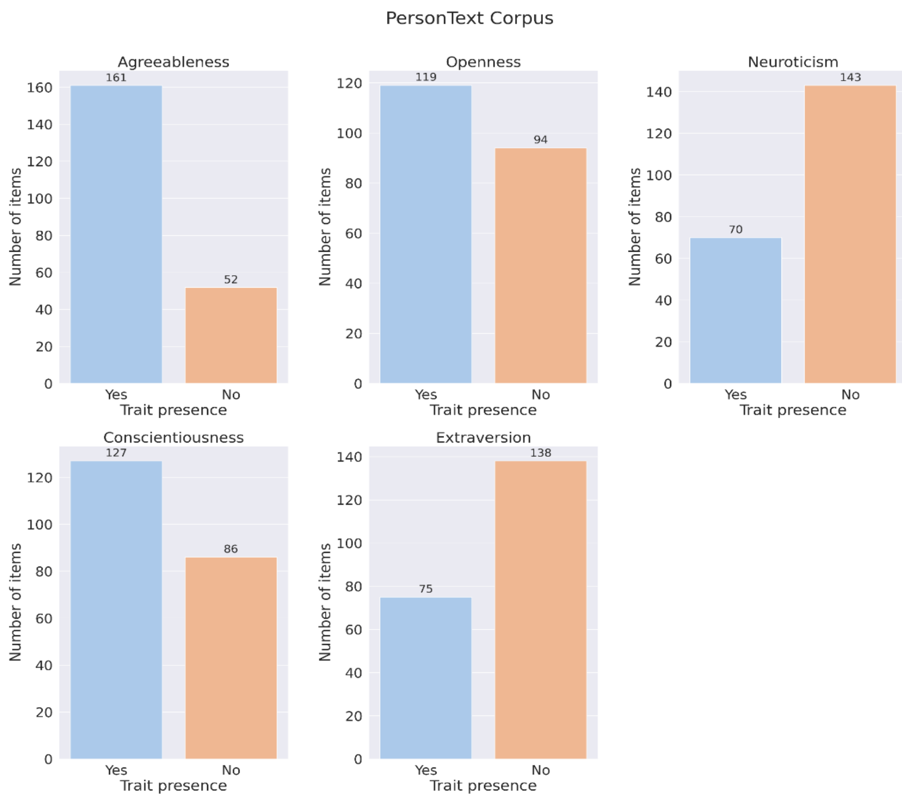

Fig. 3. PersonText corpus labels distribution

percentage of true positive predictions among the total resulting from adding the true positive predictions plus the false positives.

Independent models were used to determine each personality trait. For each personality trait, different configurations of machine learning and deep learning models were tested. Each type of model was configured with the same structure, i.e., the same configuration of hyperparameters in terms of number of epochs, layers, neurons, optimizer, activation, and loss functions, among others.

Table 3 shows and compares the accuracy results obtained by training different classifiers on the PersonText corpus, as well as two of the main corpora mentioned in the related work section (myPersonality and essays). Likewise, in Table 4, the values obtained with the precision metric are shown.

4.2 Discussion

Regarding the automatic personality recognition models assessed, it can be observed that in the case of the PersonText corpus, the most optimal results were obtained with the basic LSTM model. In both accuracy and precision, values of more than 70% were obtained, and in the five personality traits, this model obtained the best values in the metric of accuracy.

For the trait Extraversion, the Bidirectional LSTM model and the CNN model also obtained the best value, and for the trait Neuroticism, the CNN model was equal to the basic LSTM model. It can also be seen that for all traits a higher accuracy was obtained with the PersonText corpus than with the myPersonality and essays corpora.

In the precision metric, the PersonText corpus produced much higher results for the traits of

Table 3. Accuracy obtained from tests with different corpora

Corpus	Classifier	Personality traits (*Accuracy*).				
		Openness	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**
PersonText	MLP	0.68519	0.64815	0.64815	0.68519	0.64815
PersonText	Basic LSTM	**0.74074**	**0.74074**	**0.70370**	**0.72222**	**0.74074**
PersonText	Bidirectional LSTM	0.68519	0.70370	**0.70370**	0.70370	0.72222
PersonText	1D CNN	**0.72222**	0.68519	**0.70370**	0.70370	**0.74074**
myPersonality [13]	MLP	**0.73201**	0.58993	0.57914	0.55036	0.59712
myPersonality [13]	Basic LSTM	0.72302	0.65647	**0.62410**	0.60432	**0.60432**
myPersonality [13]	Bidirectional LSTM	0.72302	**0.66727**	**0.62410**	**0.60971**	0.60252
myPersonality [13]	1D CNN	0.72302	0.55935	0.57914	0.49101	0.59712
essays [8]	MLP	0.54943	0.48298	**0.54295**	0.53160	0.52836
essays [8]	Basic LSTM	**0.59968**	0.52998	0.52026	**0.55105**	**0.54781**
essays [8]	Bidirectional LSTM	0.59806	**0.53485**	0.52674	0.53485	0.53971
essays [8]	1D CNN	0.50405	0.48622	0.51540	0.50729	0.47164

agreeableness and neuroticism than for the other traits. For the trait of agreeableness, the basic LSTM model obtained a value of 0.85577, and for the trait of neuroticism, the 1D CNN model obtained a value of 0.85714.

5 Conclusions and Future Work

This research work presents the construction of a corpus in Spanish from texts extracted from videos obtained from the PersonApp data collection platform. These texts were related to personality scores for the traits of the Big-Five model.

To confirm their usefulness, automatic recognition models were developed using machine learning and deep learning techniques and results were obtained that allow us to answer our research question by concluding that it is possible to build machine and deep learning models that can obtain an acceptable performance when applied to a corpus in Spanish generated from textual transcripts of videos and their corresponding labels to personality traits based on a standardized personality test.

Likewise, it is concluded that the results obtained for the PersonText corpus with the models used were superior to those obtained for the other corpora reviewed.

The main contribution of our research is the methodology used for the creation of the PersonText corpus oriented to personality recognition in Spanish and the corpus itself. Although other corpora exist [5, 8, 11, 13], they are generally in languages other than Spanish. In the course of developing this research, we have discovered that the construction of a corpus is a significant challenge.

We consider the PersonText corpus to be a significant contribution to the field, as our corpus is oriented toward personality detection and the textual information it contains is derived from video recordings rather than social networks.

Table 4. Precision obtained from tests with different corpora

Corpus	Classifier	Personality traits (*Precision*).				
		Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
PersonText	MLP	0.67143	0.32407	0.82075	0.34259	0.32407
PersonText	Basic LSTM	**0.72931**	**0.71542**	0.70227	**0.85577**	0.75556
PersonText	Bidirectional LSTM	0.69091	0.70821	**0.72283**	0.68627	0.69444
PersonText	1D CNN	0.71841	0.65227	0.68393	0.84906	**0.85714**
myPersonality [13]	MLP	**0.79378**	0.58077	0.28957	0.55083	0.29856
myPersonality [13]	Basic LSTM	0.36151	0.65106	**0.61460**	0.60423	**0.71758**
myPersonality [13]	Bidirectional LSTM	0.36151	**0.66259**	0.61047	**0.61249**	0.57491
myPersonality [13]	1D CNN	0.36151	0.27968	0.28957	0.24550	0.29856
essays [8]	MLP	0.55019	0.47864	**0.54871**	0.53372	0.26418
essays [8]	Basic LSTM	**0.60651**	0.52834	0.51999	**0.55098**	**0.56218**
essays [8]	Bidirectional LSTM	0.59806	**0.53367**	0.52603	0.53803	0.54223
essays [8]	1D CNN	0.25203	0.24311	0.25770	0.25365	0.23582

The creation of the PersonText corpus offers an alternative for researchers wishing to tackle the task of text-based automatic personality recognition with a focus on the Spanish language. This dataset was generated from the PersonApp platform which contains information and results of standardized personality tests based on the Big-Five model.

Once the corpus was created, it was processed and subsequently tested to verify the research question. The corpus was evaluated using different neural network models to classify personality traits. Among the models, neural networks such as MLP, LSTM, Bidirectional LSTM and 1D CNN were tested, and their configuration and architecture were defined.

Five neural networks from each model were used to recognize each personality trait (Agreeableness, Openness, Conscientiousness, Extraversion and Neuroticism). The best result was 74% accuracy, which is a satisfactory result, since it is superior to those obtained by some state-of-the-art works [19] and comparable to other state-of-the-art works [21] reviewed in the related works section.

A limitation in the Spanish language is that there are not many works related to the construction of corpora oriented to automatic personality recognition based on Spanish text, since the existing data sets contain, in general, data extracted from social networks such as Facebook or Twitter in a language other than Spanish, so this research provides a corpus that can be used in future research by the Spanish-speaking scientific community.

As future work, it is proposed to increase the size of the corpus with the support of a group of expert psychologists. It is recommended to improve the automatic recognition models by exploring strategies other than those analyzed in this research. These include the use of class balancing techniques or data augmentation, which may help to improve the accuracy and precision results of the automatic recognition models. Additionally, the optimization of hyperparameters to improve the analyzed models should be explored.

6 Data Availability

The corpus is available for download[1].

Acknowledgments

The authors would like to thank the Instituto Tecnológico de Culiacán for all the support given to this research.

References

1. **Funder, D. C. (2013).** The personality puzzle (6th edition). New York: W. W. Norton & Co. (either in hardback: ISBN: 978-0-393-91311-8, or in paperback: ISBN: 978-0-393-12441-5).
2. **Myers, I. B., Myers, P. B. (2010).** Gifts differing: Understanding personality type. Nicholas Brealey.
3. **Goldberg, L. R. (1992).** The development of markers for the Big-Five factor structure. Psychological Assessment, Vol. 4, No. 1, pp. 26–42.
4. **Stillwell, D. J., Kosinski, M. (2004).** myPersonality project: Example of successful utilization of online social networks for large-scale social research. American Psychologist, Vol. 59, No. 2, pp. 93–104.
5. **Ramos, R., Neto, G., Silva, B., Monteiro, D., Paraboni, I., Dias, R. (2018).** Building a corpus for personality-dependent natural language understanding and generation. Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC'18.
6. **Bátiz-Beltrán, V. M., Zatarain-Cabada, R., Barrón-Estrada, M. L., Cárdenas-López, H. M., Escalante, H. J. (2022).** A multiplatform application for automatic recognition of personality traits for learning environments. International Conference on Advanced Learning Technologies (ICALT), Vol. 2022, pp. 49-50, DOI: 10.1109/ICALT55010.2022.00022.
7. **IPIP (2023).** International personality item pool: A scientific collaboratory for the development of advanced measures of personality traits and other individual differences. Recover from http://ipip.ori.org/.
8. **Pennebaker, J. W., King, L. A. (1999).** Linguistic styles: language use as an individual difference. Journal of personality and social psychology, Vol. 77, No. 6, 1296–1312. DOI: 10.1037/0022-3514.77.6.1296.
9. **John, O. P., Donahue, E. M., Kentle, R. L. (1991).** The big-five inventory-version 4a and 54. Berkeley, CA: Berkeley Institute of Personality and Social Research, University of California.
10. **Kim, L., Walter, D. (2008)**. Personae: A corpus for author and personality prediction from text. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, European Language Resources Association (ELRA).
11. **Rangel-Pardo, F. M., Celli, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W. (2015).** Overview of the 3rd author profiling task at PAN 2015. CLEF 2015 Evaluation Labs and Workshop Working Notes Papers, pp. 1– 8.
12. **Rammstedt, B., John, O. P. (2007).** Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. Journal of Research in Personality, Vol. 41, No. 1, pp. 203–212. DOI: 10.1016/j.jrp.2006.02.001.
13. **Celli, F., Pianesi, F., Stillwell, D., Kosinski, M. (2013).** Workshop on computational

[1] https://catalabs.mx/datasets/persontext/

ISSN 2007-9737

personality recognition: Shared task. AAAI Workshop - Technical Report.

14. **Ramírez-de-la-Rosa, A. G., Villatoro-Tello, E., Jiménez-Salazar, H. (2018).** HWxPI: A multimodal spanish corpus for personality identification. Latin American and Iberian Languages Open Corpora Forum.
15. **Gosling, S. D., Rentfrow, P. J., Swann, W. B. (2003).** A very brief measure of the big-five personality domains. Journal of Research in Personality, Vol. 37, No. 6, pp. 504–528. DOI: 10.1016/S0092-6566(03)00046-1.
16. **Xue, D., Wu, L., Hong, Z., Guo, S., Gao, L., Wu, Z., Zhong, X., Sun, J. (2018).** Deep learning-based personality recognition from text posts of online social networks. Appl Intell Vol. 48, pp. 4232–4246. DOI: 10.1007/s10489-018-1212-4.
17. **Rissola, E. A., Bahrainian, S. A., Crestani, F. (2019).** Personality recognition in conversations using capsule neural networks. IEEE/WIC/ACM International Conference on Web Intelligence pp. 180–187.
18. **Tadesse, M., Lin, H., Xu, B., Yang, L. (2018).** Personality predictions based on user behavior on the facebook social media platform. IEEE Access, pp. 1–1. DOI: 10.1109/ACCESS.2018.2876502.
19. **Majumder, N., Poria, S., Gelbukh, A., Cambria, E. (2017).** Deep learning-based document modeling for personality detection from text. IEEE Intelligent Systems, Vol. 32, pp. 74–79. DOI: 10.1109/MIS.2017.23.
20. **Mairesse, F., Walker, M., Mehl, M., Moore, Roger. (2007).** Using linguistic cues for the automatic recognition of personality in conversation and text. Journal Artificial Intelligence. Research, Vol. 30. pp. 457–500. DOI: 10.1613/jair.2349.
21. **Tandera, T., Hendro, Suhartono, D., Wongso, R., Prasetio, Y. (2017).** Personality prediction system from facebook users. Procedia Computer Science, Vol. 116, pp. 604–611. DOI: 10.1016/j.procs.2017.10.016.
22. **Liu, F., Perez, J., Nowson, S. (2016).** A recurrent and compositional model for personality trait recognition from short texts. Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES), The COLING 2016 Organizing Committee, Osaka, Japan, pp. 20–29.
23. **Zhang, A. (2015).** Speech recognition (Version 2.1) [Software]. Available from https://github.com/Uberi/speech_recognition#readme.

Article received on 26/05/2023; accepted on 07/06/2024.
[*]Corresponding author is María Lucía Barrón-Estrada.