TxPI-u: A Resource for Personality Identification of Undergraduates[*]

Gabriela Ramírez-de-la-Rosa[†1], Esaú Villatoro-Tello[1], and Héctor Jiménez-Salazar[1]

[1]Language and Reasoning Research Group, Information Technologies Department, Universidad Autónoma Metropolitana (UAM) Unidad Cuajimalpa, México
{gramirez,evillatoro,hjimenez}@correo.cua.uam.mx

May 2018

Abstract

Resources such as labeled corpora are necessary to train automatic models within the natural language processing (NLP) field. Historically, a large number of resources regarding a broad number of problems are available mostly in English. One of such problems is known as Personality Identification where based on a psychological model (e.g. The Big Five Model), the goal is to find the traits of a subject's personality given, for instance, a text written by the same subject. In this paper we introduce a new corpus in Spanish called Texts for Personality Identification (TxPI). This corpus will help to develop models to automatically assign a personality trait to an author of a text document. Our corpus, TxPI-u, contains information of 416 Mexican undergraduate students with some demographics information such as, age, gender, and the academic program they are enrolled. Finally, as an additional contribution, we present a set of baselines to provide a comparison scheme for further research.

Keywords: Language resource; Personality identification; Author profiling; Natural language processing

1 Introduction

There is a growing interest in the computer science community on studying individual's personality. This new interest, mainly among natural language processing community

---

[*]Preprint of [24]. The final publication is available at IOS Press through `https://content.iospress.com/articles/journal-of-intelligent-and-fuzzy-systems/ifs169484`

[†]Corresponding author.

1

is due to the fact that through traditional techniques developed by psychologists, identification of one's personality has been proved efficient for predicting thought patterns, emotions and behaviour [7]. Particularly, knowing this kind of information from an individual has been useful to detect her/his well-being as well as it is been helpful in the study of her/his mental health. For instance, in the past, knowing the personality of a person allowed physicians to prevent mental disorders or mental conditions [19].

Automatically identifying a person's personality is a relevant task in several areas of Computational Sciences. As an example, in the Human-Computer Interaction field, on one hand, knowing the personality of an user can help to improve its experience with the system; on another hand, providing with a compatible personality to the system itself may result in a more natural interaction with the final user [3]. For instance, in video games, the notion of a character's personality has been a key factor to improve characters' credibility [1]; in education, automated tutors could be more effective in reaching students if the tutor adapts to the student's personality [13].

In order to build systems like the ones in previous examples, it is necessary to have resources, i.e. labelled corpora, to be used in building automatic systems to effectively identify a person's personality. Historically, a large number of resources regarding a broad number of problems are available mostly in English. Particularly, for personality identification task there are very few resources available (English included) in comparison to other older problems.

This lack of resources difficult the development of such systems. In some cases, researchers have to build their own dataset which is an expensive task, in time, effort and money. Whilst there have been some attempts to build public available resources, to the best of our knowledge, there are no such resources for Spanish.

In order to overcome this obstacle and to help advance forward the research in this growing area we introduce a resource of Spanish writing texts of 416 Mexican undergraduates students. Each text is accompanied with the personality traits obtained with a psychological instrument called TIPI (Ten Item Personality Inventory); additionally, each text is also accompanied with the gender and age of the author. This extra information can be also useful for researchers investigating problems such as author profiling in Spanish texts.

The rest of this paper is organized as follows. First, in Section 2 the Big Five Personality Model (BF) is described to illustrate what is been measured to each individual participating in our study. Section 3 shows some related work to the current datasets for personality identification. Section 4 presents step by step how the corpus proposed was compiled, as well as some general information about it. After that, a statistical analysis of our corpus can be found in Section 5. Once we know the gist of our corpus, Section 6 shows some similarities between our introduced corpus against a bigger corpus of essays written in English, this English corpus is one of the most used in the NLP community. As an additional contribution, in Section 7 a set of baselines is presented using TxPI-u corpus. And finally, in Section 8 conclusions and perspectives are presented.

2

2 Big five personality model

The individual's personality is determined by her or his stable patterns of behavior shown in any particular situation. In other words, the personality is defined by those characteristics that do not change, and that are independent of the situation in which a person is involved [27].

Goldberg established that psychological models based on traits are more efficient for measuring aspects in the life's subject [9]. Such personality traits are internal dispositions that exhibit processes such as thinking, feeling or acting in specific situations resulting in the same result [28].

The dominant model based on traits is known as the Big Five Model (BF) or Five-Factor Model (FFM) [16]. This model proposes five traits with two poles each, positive and negative. The descriptions of these traits are as follows:

- *Extroversion* is associated with energy, positive emotions, assertively, sociability and expressively; its negative pole is *introversion*.

- *Emotional stability* is associated with controlling impulses, its negative pole is *neuroticism* which is the tendency to experience unpleasant emotions such as angry, anxiety, depression or vulnerability.

- *Agreeableness* refers to the tendency to be understanding and cooperative. Its negative pole refers to distrust and apathy towards others.

- *Conscientiousness* is the tendency to show auto-discipline, to act in a loyal way, to reach goals and to plan, to be organize and trustworthy. Its negative pole refers to spontaneous behaviors.

- *Openness to experience* is associated with appreciation of unusual ideas, and with imaginative and curious minds. The negative pole of this trait is associated with being unimaginative and inflexible to change.

Traditionally, to identify to what extend each trait is present in one individual, psychologists have developed standard questionnaires. The more frequently used questionnaires are: NEO-Personality-Inventory Revised (NEO-PI-R with 240 questions) [5] and the Big-Five Inventory (BFI with 44 questions) [12]. In this study we used the Ten Item Personality Inventory (TIPI with only 10 questions) [10].

3 Related work

According to Pennebaker, language is a good indicator of our personality, that is because through language we can express our way of thinking and feeling [22]. Consequently, there is a great amount of studies focusing on the analysis of expressions of language, such as those present in texts produced by a person. One of the first works in this area was based entirely on identifying types of words used in such texts [2]. Following this line of research, several other studies have conducted analysis of written texts from blogs or essays [15, 18, 11]; or in social media [6, 4, 20].

While some researchers have gathered their own resources to conduct their investigations, there are few main resources broadly used: i) a corpus collected by Pennebaker [23] and Mairesse [15] (*Essays corpus*) that consists of 2479 essays from psychology students, ii) *myPersonality corpus*, a collection of Facebook's posts [14], and iii) the *PAN-AP-15 corpus* developed in the framework of the PAN 2015[1] author profiling shared task [25], where Twitter data is provided in several languages including Spanish. These freely available datasets are different in their nature, whilst *Essays corpus* has long texts and has been gathered through several years, *myPersonality* and the *PAN-AP-15* corpora are examples of a massive short-texts data gathered from social media domains.

Although the *PAN-AP-15 corpus* has small partition of tweets in Spanish, our corpus is more directly related to the Essays corpus. As mentioned before, our goal is to contribute in providing resources for promoting studies of personality identification in the Spanish language. In addition, we want to provide a reference study on the performance of existing methods and algorithms in solving the posed task using our corpus. In Section 6 a more detailed comparison between Essay corpus and TxPI-u corpus is presented.

4   Making the TxPI-u corpus

The TxPI-u (Text for Personality Identification of Undergraduates) corpus is a resource that can be used for building automatic systems for personality identification task. This corpus consists of texts written in Spanish from undergraduates Mexican students. In the following we describe the methodology employed to assemble such corpus.

4.1   The sample

Every year the Autonomous Metropolitan University campus Cuajimalpa (Universidad Autónoma Metropolitana Unidad Cuajimalpa) receives undergraduate students. During 2016, the University received near 600 students for 10 different academic programs.

During the fourth week of classes we attended to the classrooms to ask for the students' participation in our study. The students were informed about the research we were conducting and 417 decided to collaborate. The distribution of participants (also referred as subjects) per academic program is shown in Table 1.

As we can see the corpus is balanced in terms of gender. There is also a representation of each field of study from social sciences to mathematics and engineering. Since the subjects were starting the undergraduate education almost everyone is between 19 and 21 years old. The complete distribution of age and gender per academic program can be seen in Figure 1.

4.2   The instrument

The participation in our study consisted in answering an instrument. The goal of such instrument was twofold. First, to determine the subjects' personality in order to label

---

[1]http://pan.webis.de/clef15/pan15-web/author-profiling.html

Table 1: Participants in the corpus divided by gender per academic program in the 2016 admission process at the University

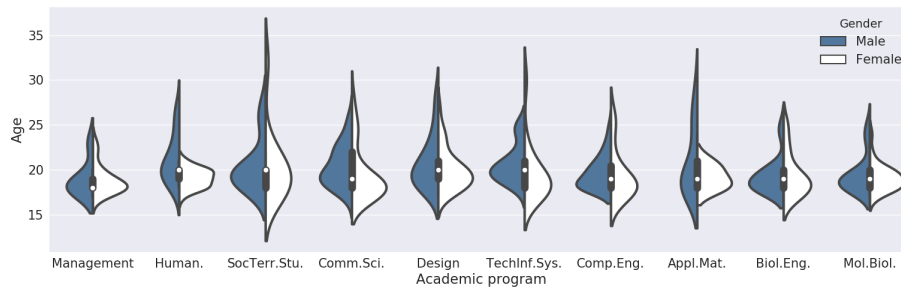| Academic program	Male	Female	Total
Management	6	17	23
Humanities	19	25	44
Social-territorial Studies	12	12	24
Communication Sciences	30	30	60
Design	15	37	52
Tech. & Information Systems	28	16	44
Computational Engineering	45	15	60
Applied Mathematics	14	5	19
Biological Engineering	20	23	43
Molecular Biology	19	29	48
Total	208	209	417

Figure 1: Distribution of subjects per age, gender and academic program. Total number of subject is 417, the data for this graph came partially from Table 1.

the data according to the Big Five model; and second, to collect a sample of written text from all the subjects regarding personal experiences.

Consequently, the instrument was designed with three parts in order to gather: i) general information, ii) answers for the personality test, and iii) a handwritten short essay.

  (i) The first section, general information, was designed to get contact information of the subject, also her or his gender, the academic program and her or his social media accounts (such as Facebook and Twitter). It is worth to mention that a small percentage of the subjects indicated a social media account; thus, such information is not included in the final corpus.

 (ii) The second section has a personality test. In order to take as few time as possible from the participants, the Ten Item Personality Inventory [10] in Spanish [26] was used.

(iii) The last part of the instrument included one instruction and a blank page. The instruction was given in Spanish and can be translated as: *Tell us about yourself, for instance, something about your family's history or an event you think was relevant in your life that comes to your mind*.

5 Description of TxPI-u corpus

5.1 The essays and its transcriptions

To ease the application of the instrument to all students, our instrument was applied on paper. During the digital transcriptions of the handwritten essays we noticed some particularities of handwriting, i.e. small modifications of words, the intent to erase a word, insertion of letters into words, or words into sentences, incorporation of emojis or drawings, as well as misspelling and syllabification. We believe that analyzing such handwriting phenomena would be useful for a better understand on how people with certain trait of personality thinks and behaves. Consequently, the TxPI-u corpus provides two version of the essays, with and without these labels. Although in a digital environment this type of phenomena would be difficult to capture, our intention in labeling such information is to analyze if there is a direct correlation among these and the subjects' personality traits.

Thus, we used seven labels, namely: <FO:well-written word> (misspelling), <D:description> (drawing), <IN> (insertion of a letter into a word), <MD> (modification of a word, that is a correction of a word), <DL> (elimination of a word), <NS> (when two words were written together; e.g. *Iam* instead of *I am*) and, SB (syllabification). An example of such tagging is given in Table 2.

We compute the Pearson correlation of the percentage of presence in each essay to analyze how these seven tags are correlated among them. Table 3 shows that the maximum value of correlations is still below 0.1. Nevertheless, the more correlated tags are *NS* and *FO* with 0.08 and also, *NS* and *MD* with 0.07.

We also measured the number of words and total vocabulary used by the analyzed subjects. Figure 2 presents the correlation between these two values and shows, as is

6

Table 2: Example of a subject hand written essay in Spanish, its manual transcription with added tags and its corresponding English translation
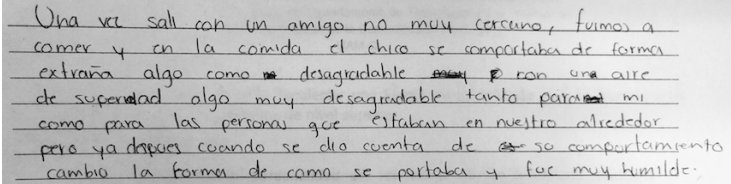

Manual transcription	Una vez sali <FO:salí> con un amigo no muy cercano, fuimos a comer y en la comida el chico se comportaba de forma extraña algo como <DL> desagradable <DL> <DL> con un <MD> aire de superioridad <MD> algo muy desagradable tanto para <DL> mi <FO:mí> como para las personas que estaban en nuestro alrededor pero ya despues <FO:después> cuando se dio cuenta de <DL> su comportamiento cambio <FO:cambió> la forma de como <FO:cómo> se portaba y fue muy humilde.
English translation	Once I went out with a friend not so close to me, we went to eat and while eating the guy was acting a little weird kind of rude as he was superior to me, it was rude for me as for the people around us but after he realized his behavior he changed the way he was acting and he was humble.

Table 3: Correlation among tags; where tags FO, D, IN, MD, DL and NS means misspelling, drawing, insertion of some letter, modification of some word, elimination of some word, do not separate two words, and syllabification, respectively

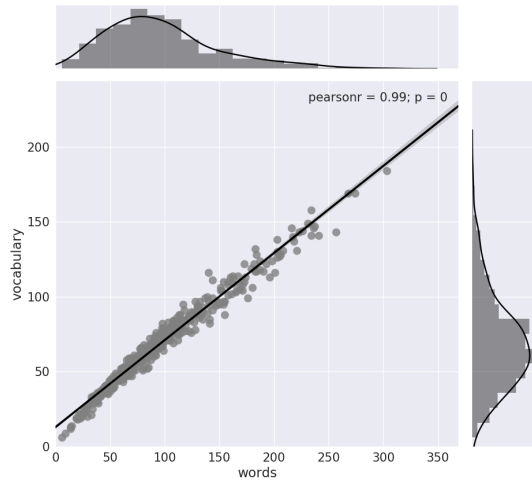|	FO	D	IN	MD	DL	NS
D	-0.04					
IN	-0.01	-0.01				
MD	0.00	0.01	0.00			
DL	0.02	-0.02	-0.03	0.05		
NS	0.08	-0.03	-0.03	0.07	0.06	
SB	-0.04	-0.02	-0.03	0.02	-0.06	0.00

Figure 2: Distribution of number of words and total vocabulary per essay and the correlation among these two variables in the TxPI-u corpus. Shown frequency distribution graph (above and right) describes one variable independently of the other.

expected, that the more written words the greater the vocabulary used. Note that to calculate this measures we did not used any kind of lemmatization tool.

5.2 Personality information

For the second part of the instrument (see Section 4.2) we registered the numeric value computed for each trait according to the answers to the Ten Item Personality Inventory (TIPI) test [10].

The TIPI test includes two questions (items) for each of the five traits of the Big Five Model. A more detailed explanation of how to compute the personality of any given answer is presented in [10]. Consequently, this test allows to have a numeric value between 1 and 7 to each trait, Figure 3 shows the distribution of this numeric value.

To have a general view of numerical values of each trait, the correlation between traits was computed with a Pearson correlation. In Table 4 all correlations are shown. It can be seen that traits more positively correlated are Emotional Stability and Agreeableness with a value of 0.34. Emotional Stability is also positively correlated to Consciousness (with 0.28). While the correlation is small, we can say that stable subjects are also, to some extend, agreeable and reasonable. Another positive correlation exists between Openness and Extroversion (0.28), indicating that subjects open to new experiences are also, to some extend, extroverts.

Then, according to Gosling's normative data for the TIPI questionnaire, the 417 subjects were classify with five traits. Each of them were labeled into four classes: high, medium high, medium low, and low. Table 5 shows the number of subjects per
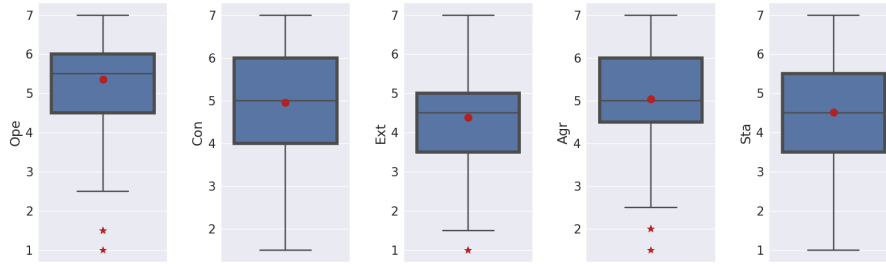
Figure 3: Distribution of numerical values of each trait according to the answers of each subject to the TIPI test. The maximum value is 7 and the minimum is 1. The media value is shown inside the box by a circle.

Table 4: Correlation values among traits according to the answers to the TIPI test. Ext, Agr, Con, Sta and Ope stand for Extroversion, Agreeableness, Consciousness, Emotional Stability, and Openness, respectively

	Ext	Agr	Con	Sta	Ope
Ext	1.00				
Agr	-0.06	1.00			
Con	0.08	0.23	1.00		
Sta	0.09	0.34	0.28	1.00	
Ope	0.27	0.15	0.16	0.07	1.00

class of each personality trait. Each trait has its own normative values; which were obtained with 1704 subjects of different ethnicities [10]. One subject was removed from the collection since her/his essay was empty; therefore, the total number of subjects in TxPI-u is 416.

5.3 Stratified partition

As described in the previous section, the TxPI-u corpus contains essays from 416 subjects, each of them presents five traits according to the Big Five Model and every subject can be found in each trait (i.e. in Table 5 the total number of subjects per trait is always 416). While this organization of TxPI-u can be useful to analyze traits independently to each other; might be a complication to analyze more than one trait in

Table 5: Number of subjects per class of each personality trait according to the normative data for the Ten-Item Personality Inventory (TIPI) given by Gosling et al. [10]

Trait	High	Medium High	Medium Low	Low
Openness	91	145	116	64
Consciousness	19	150	138	109
Extroversion	72	137	169	38
Agreeableness	60	115	151	90
Emotional Stability	34	151	151	80

Table 6: Number of subjects by gender and number of subjects by class, per trait, in the stratified partition

	Gender		Classes		
Trait	Male	Female	High	Low	Total
Openness	18	14	16	16	32
Consciousness	11	17	3	25	28
Extroversion	15	10	17	8	25
Agreeableness	12	20	10	22	32
Emot. Stability	7	7	6	8	14
Control	44	39	-	-	83
Total	107	107	52	79	214

Table 7: Percentage of tags present in essays of the stratified partition of TxPI-u corpus. The tags FO, D, IN, MD, DL and NS means misspelling, drawing, insertion of some letter, modification of some word, elimination of some word, do not separate two words, and syllabification, respectively

Tag	mean	std	min	max
FO	1.46	1.71	0	9.68
D	0.01	0.14	0	1.92
IN	0.01	0.09	0	1.22
MD	0.99	1.99	0	20.64
DL	0.22	0.72	0	6.33
NS	0.42	0.96	0	6.04
SB	0.14	0.58	0	5.56

combination. Therefore, we decided to make a stratified partition where each trait has a set of representative examples from the positive (high) and negative (low) pole; as well as a set of examples for control purposes (control sample).

In the stratified partition of the TxPI-u, the control sample contains all the subjects with classes "medium high" or "medium low" for every trait. In other words, all subjects in the control group do not have any predominant traits ("high" or "low"). Hence, there is only one control group.

Subjects with representative traits, i.e. subjects labeled with classes "high" or "low" in only *one* trait are selected for the stratified partition. This idea of stratified corpus has been done before by Oberlander and Gill, as they stated, a three-way stratified corpus allows to analyze features along a unique dimension [17].

Table 6 shows the number of male and female per trait and the number of subjects in the control group of the resulting stratified partition. As we can see, the stratified corpus is smaller (almost half of the complete corpus), but it allows to perform a more fine analysis of the differences between traits.

Regarding to handwritten phenomena, Table 7 shows the percentages of use of the seven tags. As can be noticed, the misspelling tag (FO) has the higher percentage. In Figure 4 a closed look of the distribution of FO is presented.

In Figure 4 is worth noticing the difference in the percentage of misspelling between classes of the same trait. For instance, for the trait extroversion there is almost the same percentage of misspelling for all subjects in the low class, while there is a big-
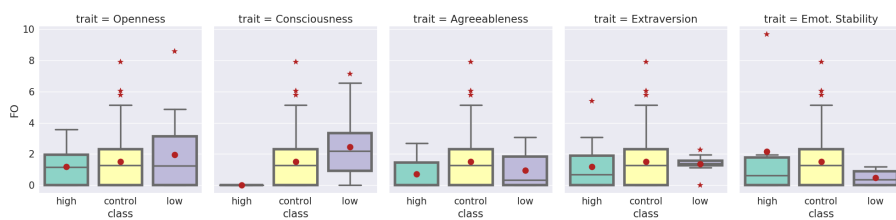
Figure 4: Distribution of the misspelling tag (FO) across traits in the stratified corpus. The red dot inside the boxes shows the medium value of each sample and the dots beyond the limit of the boxes denote outliers.

ger distribution of percentage of misspelling for subjects in the high class. That could be an indication of possible correlation of misspelling with a subject's predominant personality.

6 Comparison of corpora for personality identification

As it was mentioned in Section 3, one corpus of English texts is the more related to TxPI-u. This corpus is a collection of essays compiled by Pennebaker and King [23] during 1997 and 1999. Most of the texts from this corpus come from psychology students (approx. 1200). This initial corpus was increased by Mairesse et al. [15] with essays of students written until 2004. The final corpus, referred as *Essays corpus* is a compilation of 2468 essays, most of them written by students. There is not information about the gender and age of participants in this corpus[2].

In order to provide some general comparison between TxPI-u and the Essays corpus we show information about the number of words and vocabulary used (see Figure 5). Despite of the sample size (2468 vs 416 of Essays and TxPI-u respectively) there is a similar correlation between the number of words and the vocabulary used for each subject. In Figure 6 a direct comparison between two variables is shown side by side. As can be seen, the texts in Essays are larger but also have more variation in the number of words used as in the vocabulary. This variation appears to a lesser extent in TxPI-u. The difference between the number of words used in both corpora could be explained in terms to the instruction given to the subjects. While in our case we asked for a personal experience, in the Essay corpus compilation, authors asked the subjects to write, for 20 uninterrupted minutes, anything that came to their minds (a complete description of the Essay corpora compilation can be found in [23]).

Additionally, Table 8 shows the number of subjects per trait, note that there is not information about the numerical values of subjects' personality traits, only the nominal class "yes" or "no" was provided (the class "yes" is similar to our class "high" and the class "no" is similar to our class "low"). With numerical values of personality's traits another normative values can be used to generate other partition of nominal classes

---

[2]In [23] there is partial information about gender and age of 1200 participants approximately.
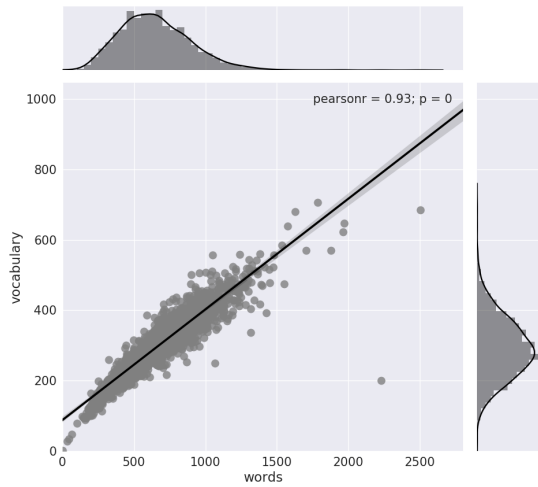
Figure 5: Distribution of number of words and total vocabulary per essay and the correlation among these two variables in the Essays corpus. Frequency distribution graph (above and right) describe one variable independently of the other.
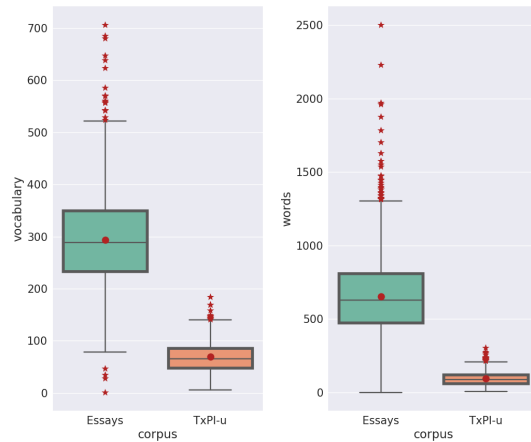

Figure 6: Comparison of number of words and total vocabulary of each texts between the Essays corpus and the TxPI-u corpus. The red dot inside the boxes shows the medium value of each sample and the dots beyond the limit of the boxes denote outliers.

12

Table 8: Number of subjects per trait in the Essays corpus. Note that labels *yes* and *no* correspond to *high* and *low*, respectively

Trait	yes	no
Openness	1271	1196
Consciousness	1254	1214
Extroversion	1277	1191
Agreeableness	1310	1158
Emot. Stability	1235	1233

or, a regression approach can be used to determined such values instead of just closed categories.

7 Text classification with TxPI-u corpus: baselines

The main goal of this section is to provide a set of baselines for the text classification task of personality identification. All the experiments reported in this section were done using the stratified partition of the TxPI-u corpus.

We provide a set of basic configuration systems, widely employed in the text classification (TC) task. Obtained results will serve for comparison purposes against future methods or future representations. The intention is to use representations such as n-grams of words, n-grams of characters and n-grams of part of speech (POS) in combination with the most common learning algorithms for text classification such as naive bayes, decision trees and support vector machines.

7.1 Evaluation metrics

The evaluation metric used was the macro-averaged F_1, also known as F-score. This measure allows to obtain confident perspectives of the system's performance, particularly for cases where classes are highly unbalanced, such as in the stratified partition of TxPI-u.

7.2 Experimental setup

Five classification problems were defined, one per trait. Each problem has three classes: high, low, and control. We represented each essay using three different type of representations: n-grams of words, n-grams of characters and n-grams of POS (Part of Speech) tags. For each type we used n-grams' sizes of 1, 2 and 3 for words and POS, and n-grams' sizes of 3, 4, 5 for characters. In addition, for each experiment we used three different classifiers: Naive Bayes, J48 and SMO[3].

A vector space model was used to represent each text; thus, for each essay we have a multi-dimensional vector. In this vector, we evaluated three different weighing schemes: boolean (the importance of each term in the vector should be 1 if the term appears in the document, and 0 otherwise), term frequency (the number of times a

---

[3]For all experiments we used Weka Tool Kit [8] with the default configuration values.

Table 9: Classification results using ten cross fold validation technique with the stratified corpus of TxPI-u. Each result corresponds to a three class classification problem. Results are given in F-score and the representation used is BOW with LIWC tags

	NB		SMO		J48	
	bool	tf	bool	tf	bool	tf
Ope	0.31	0.33	**0.35**	0.26	0.34	0.34
Con	0.30	**0.36**	0.29	0.33	0.26	**0.36**
Ext	0.31	0.34	**0.35**	0.34	0.34	0.33
Agr	0.34	0.40	0.31	0.27	**0.43**	0.36
Sta	0.35	**0.40**	0.27	0.30	0.31	0.37

term appears in the document), and tf-idf (the importance of a term given by the term frequency and the inverse document frequency).

7.3 Results

Altogether, we performed 405 experiments (5 traits, 9 representations, 3 weights schemes, and 3 learning algorithms) and for all of them we used 10 fold cross validation to evaluate each three-class classifier. For all results we calculate the F-score that can be seen in Table 10. Note that in the results' table only boolean (bool) and term frequency (tf) as weighting schema is shown, that is because tf-idf (term frequency - inverse document frequency) performs similar to tf, therefore, we only show tf results.

Additionally, a set of experiments were done using a Bag of Words representation with the words categories presented in the Spanish dictionary (version 2007) of LIWC (Linguistic Inquiry and Word Count) [21]. The results of this experiment also reported using the F-score metric; see Table 9.

Overall, the performance of a three-class problem is not superior to 0.49 of F-score which illustrates the difficulty of this problem. We believe that by means of novel representation schemes or new learning methods, obtained results can be significantly improved.

Despite the low performances, we can have some interesting insights about the representations used. For instance, the representation based on POS tags is the best across three traits: Openness (f-score of 0.49 with uni-grams), Consciousness (f-score of 0.39 with bi-grams) and Emotional Stability (f-score of 0.40 with uni-grams); while 5-grams of characters performed better for Extroversion and Agreeableness (0.45 of f-score for both cases).

Regarding the representation using word categories of LIWC, the performance of classification was not better that using an open-vocabulary approach (as in the experiments reported in Table 10). It is clear that these results are not definitive, and there still is an important room for improvement, e.g. a representation that combine open-vocabulary with word categories. Even more, a deeper analysis can be performed to better understand the difficulty of the personality identification problem.

Table 10: Classification results using a ten fold cross validation technique with the stratified corpus of TxPI-u. Each result corresponds to a three-class classification problem. Results are given in F-score

Words

	1-gram						2-gram						3-gram					
	NB		SMO		J48		NB		SMO		J48		NB		SMO		J48	
	bool	tf	bool	tf	bool	tf	bool	tf	bool	tf	bool	tf	bool	tf	bool	tf	bool	tf
Ope	0.28	0.34	0.28	0.28	0.35	0.36	0.28	0.33	0.28	0.28	**0.38**	0.32	0.28	0.27	0.28	0.28	0.28	0.27
Con	0.27	0.32	0.30	0.28	0.30	0.36	0.27	0.35	0.28	0.28	**0.38**	0.32	0.29	0.33	0.29	0.29	0.28	0.32
Ext	0.29	0.31	0.29	0.29	**0.36**	**0.36**	0.29	0.27	0.29	0.29	0.26	**0.36**	0.29	0.34	0.29	0.29	0.29	0.29
Agr	0.35	**0.45**	0.35	0.31	0.31	0.35	0.28	0.26	0.28	0.28	0.30	0.33	0.28	0.24	0.28	0.28	0.27	0.27
Sta	0.31	0.39	0.31	0.31	0.35	0.39	0.31	0.36	0.31	0.31	0.31	**0.41**	0.31	0.37	0.31	0.31	0.31	0.31

Part of Speech

	1-gram						2-gram						3-gram					
	NB		SMO		J48		NB		SMO		J48		NB		SMO		J48	
	bool	tf	bool	tf	bool	tf	bool	tf	bool	tf	bool	tf	bool	tf	bool	tf	bool	tf
Ope	0.41	0.40	0.37	**0.49**	0.38	0.42	0.29	0.30	0.30	0.30	0.34	0.44	0.32	0.34	0.28	0.28	0.24	0.28
Con	0.36	0.35	0.42	0.34	0.31	0.30	0.28	**0.39**	0.30	0.30	0.28	0.33	0.30	0.37	0.29	0.28	0.29	0.27
Ext	**0.39**	0.32	0.37	0.28	0.28	0.34	0.28	0.28	0.35	0.31	0.30	0.27	0.29	0.27	0.29	0.29	0.31	0.25
Agr	0.34	0.33	**0.41**	0.37	0.29	0.32	0.34	0.29	0.33	0.29	0.31	0.30	0.34	0.29	0.28	0.28	0.26	0.29
Sta	0.33	0.28	**0.46**	0.31	0.31	0.29	0.31	0.31	0.38	0.31	0.41	0.41	0.31	0.31	0.31	0.31	0.37	0.29

Characters

	3-gram						4-gram						5-gram					
	NB		SMO		J48		NB		SMO		J48		NB		SMO		J48	
	bool	tf	bool	tf	bool	tf	bool	tf	bool	tf	bool	tf	bool	tf	bool	tf	bool	tf
Ope	0.31	0.28	0.28	0.27	0.37	0.26	0.28	0.28	0.28	0.28	0.28	0.33	0.27	0.27	0.28	0.28	**0.38**	**0.38**
Con	0.34	0.28	0.30	0.31	0.37	0.36	0.35	0.28	0.34	0.30	**0.38**	0.33	0.32	0.33	0.31	0.31	0.29	0.31
Ext	0.35	0.29	0.36	0.33	0.33	0.33	0.32	0.28	0.29	0.29	0.30	0.29	0.32	0.29	0.29	0.29	**0.45**	0.41
Agr	0.36	0.27	0.35	0.41	0.38	0.35	0.37	0.28	0.33	0.33	0.26	0.37	0.39	0.37	0.28	0.28	**0.45**	0.43
Sta	0.31	0.31	0.31	0.31	0.29	0.35	0.31	0.31	0.31	0.31	0.27	**0.44**	0.31	0.31	0.31	0.31	0.31	0.30

8 Conclusions

In this paper we introduced a corpus of Spanish texts annotated with personality information, age and gender for each author of those texts. The corpus, named TxPI-u, Text for Personality Identification of undergraduate, is a collection of 416 short essays of undergraduate Mexican students. We have described TxPI-u corpus in terms of number of words, vocabulary and distribution frequencies among personality traits, particularly within the Big Five personality model.

From the texts in the 416 essays, we manually labelled some handwriting phenomena, such as, modification of a word, insertion of some letter, the use of emojis or drawings, etc. We also labelled all the misspelling words found in those essays as well as syllabification acts. In this direction, we found that there is no clear correlation among the seven labels used. Nevertheless, this manually labelling of handwritten phenomena is not found, to the best of our knowledge, in any other corpus for personality identification. A work perspective is a correlation study of the presence of each label and a personality trait, one intuition is that misspelling can be associated with Consciousness in the low class.

In order to allow a fine analysis of personality traits, we created a stratified partition from TxPI-u corpus. The resulting partition contains a total of 214 subjects. In this direction, a work perspective is to study a binary classification problem, with one class for the pole of interest (either high or low) and the other class with only the control class. In this configuration only important markers for a class of a given trait will be analyzed in isolation, thus correlations can be found without introducing noise. A different, yet interesting line of work, would be to face the personality identification

problem as a regression problem, i.e., identify the numeric values of each trait.

As an additional contribution of this work, we describe a the set of baselines to provide a comparison point for further research in author profiling, specifically for personality identification. For these experiments, we used as main form of representation n-grams of: words, characters and part of speech, as well as, word categories provided by the LIWC Spanish dictionary. Tackling both, closed-vocabulary approach and open-vocabulary approach.

Finally, we believe that this resource, carefully gathered, represents an important contribution to the community doing research in the area of personality identification, and in general for the author profiling research field. With the extra information attached to the essays, automatic models can be proposed for identifying gender, age, and academic program election. In addition, our built corpus has a multimodal applicability, since it could be interesting to analyze the handwritten phenomena using novel computer vision strategies. All these characteristics make the TxPIu corpus a more challenging dataset. The complete corpus as well as the instrument used for collected it can be found at `lyr.cua.uam.mx/resources/personality/TxPIu/`.

Acknowledgments

References

[1] Elisabeth André, Martin Klesen, Patrick Gebhard, Steve Allen, and Thomas Rist. *Integrating Models of Personality and Emotions into Lifelike Characters*, pages 150–165. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000.

[2] Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James W. Pennebaker. Lexical predictors of personality type. In *In Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*, 2005.

[3] Timothy W. Bickmore and Rosalind W. Picard. Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput.-Hum. Interact.*, 12(2):293–327, June 2005.

[4] Fabio Celli and Luca Polonio. Relationships between personality and interactions in facebook. In *Social Networking: Recent Trends, Emerging Issues and Future Outlook*, pages 41–54. Nova Science Publishers, Inc., 2013.

[5] Paul T. Costa and Robert R. McCrea. *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI)*. Psychological Assessment Resources, Odessa, Fla. P.O. Box 998, Odessa 33556, 1992.

[6] Golnoosh Farnadi, Geetha Sitaraman, Shanu Sushmita, Fabio Celli, Michal Kosinski, David Stillwell, Sergio Davalos, Marie-Francine Moens, and Martine De Cock. Computational personality recognition in social media. *User Modeling and User-Adapted Interaction*, 26(2):109–142, Jun 2016.

[7] David C. Funder. Personality. *Annual Review of Psychology*, 52(1):197–221, 2001.

[8] Stephen R. Garner. Weka: The waikato environment for knowledge analysis. In *Proceeding of The New Zealand Computer Science Research Students Conference*, pages 57–64, 1995.

[9] Lewis R. Goldberg. The structure of phenotypic personality traits. *American Psychologist*, 48(1):26–34, 1993.

[10] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6):504 – 528, 2003.

[11] Francisco Iacobelli, Alastair J. Gill, Scott Nowson, and Jon Oberlander. *Large Scale Personality Classification of Bloggers*, pages 568–577. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[12] Oliver P John, Eileen M Donahue, and Robert L Kentle. The big five inventory—versions 4a and 54. University of California, Berkeley, Institute of Personality and Social Research, 1991.

[13] Meera Komarraju and Steven J. Karau. The relationship between the big five personality traits and academic motivation. *Personality and Individual Differences*, 39(3):557 – 567, 2005.

[14] Michal Kosinski, Sandra C. Matz, Samuel D. Gosling, Vesselin Popov, and David Stillwell. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *Journal of Personality and Social Psychology*, 70(6):543–556, 2015.

[15] François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research (JAIR*, pages 457–500, 2007.

[16] R. R. McCrae. Cross-cultural research on the five-factor model of personality. *Online Readings in Psychology and Culture*, 4(4), 2002.

[17] Jon Oberlander and Alastair J. Gill. Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes*, 42(3):239–270, 2006.

[18] Jon Oberlander and Scott Nowson. Whose thumb is it anyway? classifying author personality from weblog text. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 627–634, Sydney, Australia, July 2006. Association for Computational Linguistics.

[19] Daniel J. Ozer and Verónica B. Martínez. Personality and the Prediction of Consequential Outcomes. *Annual Review of Psychology*, 57(1):401–421, 2006.

[20] Gregory Park, H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Michal Kosinski, David J. Stillwell, Lyle H. Ungar, and Martin E. P. Seligman. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108(6):934–952, 2015.

[21] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth. The Development and Psychometric Properties of LIWC2007. This article is published by LIWC Inc, Austin, Texas 78703 USA in conjunction with the LIWC2007 software program.

[22] James W. Pennebaker. *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury Press, New York, 1st edition, 2011.

[23] James W. Pennebaker and Laura A. King. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296–1312, 12 1999.

[24] Gabriela Ramírez-de-la Rosa, Esaú Villatoro-Tello, and Jiménez-Salazar Héctor. Txpi-u: A resource for personality identification of undergraduates. *Journal of Intelligent & Fuzzy Systems*, 34(5):2991–3001, May 2018.

[25] Francisco Manuel Rangel Pardo, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. Overview of the 3rd author profiling task at pan 2015. In *CLEF 2015 Evaluation Labs and Workshop Working Notes Papers*, pages 1–8, 2015.

[26] Vanessa Renau, Ursula Oberst, Sam Gosling, Jordi Rusiñol, and Ander Chamarro. Translation and validation of the ten-item-personality inventory into spanish and catalan. *Aloma: Revista de Psicologia, Ciències de l'Educació i de l'Esport*, 31(2), 2013.

[27] Alessandro Vinciarelli and Gelareh Mohammadi. A survey of personality computing. *IEEE Transaction on Affective Computing*, 2014.

[28] Cornelia Wrzus and Matthias R. Mehl. Lab and/or field? measuring personality processes and their social consequences. *European Journal of Personality*, 29(2):250–271, mar 2015.