A System for Personality and Happiness Detection

Yago Saez[1], Carlos Navarro[1], Asuncion Mochon[2] and Pedro Isasi[1]

[1]*University Carlos III of Madrid, Computer Science Department, Madrid, Spain*
[2]*Applied Economics Department, UNED, Madrid, Spain*

Abstract — **This work proposes a platform for estimating personality and happiness. Starting from Eysenck's theory about human's personality, authors seek to provide a platform for collecting text messages from social media (*Whatsapp*), and classifying them into different personality categories. Although there is not a clear link between personality features and happiness, some correlations between them could be found in the future. In this work, we describe the platform developed, and as a proof of concept, we have used different sources of messages to see if common machine learning algorithms can be used for classifying different personality features and happiness.**

Keywords — **personality detection, Android OS, happiness, written text, machine learning, classifying algorithms**

I. Introduction

Since Hans Jürgen Eysenck in 1947 defined the pillars, or traits, that form personality [1], numerous studies have been conducted and many works have been written about the subject, see Section II. These works have supported his theory of individual differences between humans with regards to personality. This theory is also known as the PEN model because of the three traits on which it is based: Psychoticism, Extroversion and Neuroticism. The theory provides a direct way to obtain a score for each component by using questionnaires, specifically the EPQ-R questionnaire. Each of the three personality traits has a biological basis, so the scores obtained for the traits represent different brain processes.

Researchers have tried to obtain information about the personality of human beings through direct means such as the EPQ-R questionnaire, but they have also used indirect methods. Because personality is considered to be stable over time and throughout different situations, specialized psychologists are able to infer the personality profile of a subject by observing the subject's behavior.

One of the sources of knowledge about the behavior of individuals is written text. According to research in this field, it is reasonable to expect that different individuals will have different ways of expressing themselves through the written word, and these differences will correspond to their individual personality profiles, as well as their moods.

This type of reports offers numerous advantages for researchers because a substantial amount of information about a subject's personality profile can be obtained without their presence or any additional specific effort on the subject's part.

A. Multidisciplinary work

Although research on personality profiling and analysis of the written word is part of psychology, collaboration with other disciplines, such as computer science, is necessary for certain purposes. Even with a solid psychological theoretical foundation, it is also necessary to be able to use quantitative methods to analyze large amounts of information. Such methods are especially applicable when analyzing large amounts of written text.

It is thus necessary to undertake this type of research with a multidisciplinary team, in which social sciences researches and computer scientists combine their knowledge to create efficient tools for the analysis of human personality. Computer science provides the tools necessary to collect, process and classify text samples of psychological interest in a systematic fashion, based on the principles of software engineering and artificial intelligence.

A tool with the aforementioned characteristics will be of great interest for the economy and human happiness. For example, if a system that could recognize the personality traits of a criminal in a matter of minutes with a high degree of confidence was available to law enforcement, a more efficient handling of critical situations could be achieved.

The remainder of this article is structured in the following manner. The following section describes the most relevant works related to this research. Section III describes the objectives and answers to common questions. Section IV depicts Eysenck`s theory of personality background. After Section IV, we describe the proposed platform, (Section V), the classifier module (Section VI) and the preliminary results (section VII). Finally, the main conclusions are presented in Section VIII.

II. State of the Art

The U.S. Army War College has shown an interest in predicting and controlling the behavior of an individual or

group of individuals based on knowledge of their personalities. They believe that a system capable of this would have important applications in State security, competition in the labor market, political elections, or simply in the acquisition of knowledge about any person whose behavior might be of interest, see [2].

To perform a strategic personality simulation, they recommend taking into account the intersection between internal and external elements as well as external situational factors and personal influences.

Professors of computer science Gill and Oberlander conducted [3] a study on the recognition of the "extroversion/introversion" personality trait based on written text. They based their work on the Eysenck model [4]. For this purpose, they asked subjects with known scores on the EPQ-R questionnaire to write two e-mails to a fictitious friend. They subsequently analyzed these e-mails with a text analysis program called LIWC (*Linguistic Inquiry and Word Count*) and with the psycho-linguistic database MRC. They generated bigram profiles according to the degree of extroversion of the subjects (high or low). The results showed differences between the two sub-types of samples. Based on these differences, it was found that extroverts use more punctuation and exclamation signs, produce texts with more words, make more references to social situations, and use a greater number of positive words. Introverts, in contrast, are more likely to use the first-person singular, express themselves using more emotionally negative words, and use more coordinating conjunctions. The researchers also made lists of frequently used bigrams for both groups.

With their results, both authors conclude that the personality dimensions have relevance and validity for working with human-computer communication and computer learning.

Young presents in 2003 a geographical profiling, which consists of the profiling of criminals based on questions such as "when" or "where," instead of based on their motivations, age, gender, or other indicators [5]. With this approach, the need to incorporate computer science into the profiling process is emphasized to analyze large databases and prevent people from overlooking important information or connections between crimes. This type of analysis becomes imperative in the case of serial killers, who may commit crimes in different states that involve victims who do not know each other. The proposal coincides with the nature of this project in that it warns about the need for interdisciplinary work and highlights the importance of computer science for the processing of data that individual psychologists would not be able to analyze manually.

In this article [6], the principle of geographic profiling is presented. Geographic profiling is an attempt to obtain a wide body of information about criminal cases to provide a general psychological description of an unknown subject (UNSUB) — a possible suspect. After going into detail about the description of geographic profiling, the author presents several programs for collecting the essential information for this purpose. First,

the Violent Criminal Apprehension Program (VICAP) is presented, which is used by the FBI to efficiently analyze the connections between existing criminal cases. Second, Kim Rosso's Criminal Geographic Targeting (CGT) is exhibit. This computer program produces a topographic map by performing many calculations that group together similar crimes, and it takes into account human movement patterns. Lastly, the Predator system, developed by Dr. Grover and M. Godwin, is described. This system uses multivariate analysis to carry out geographic profiling and produces a 3D, color-coded map to classify different areas according to the probability that the perpetrator lives or operates in them.

The word done by F. Mairesse and M. Walker may be considered to be the most important antecedent of the System for Personality Detection (SPD) project [7]. The researchers attempted to automatically identify personalities based on pieces of recorded conversations. Their personality analysis was based on the Five Factor Model (see [8]), which, is closely related to the personality traits of the PEN model used in the present project. In addition to confirming previous studies, the authors reached conclusions about personality. For example, they found that correlations between linguistic indicators and personality traits are higher in informal spoken dialog; this conclusion has stimulated the use of informal language in SPD. They also concluded that the most complex trait to analyze is "neuroticism," whereas "agreeableness" and "conscientiousness" provide the best results. Prosodic indicators were found to be the most accurate predictors for "extroversion." Finally, they concluded that their hypothesis, which proposes that it is possible to automatically detect personality through language, is confirmed, and they find that their procedure is applicable to a variety of fields.

The work of T. Polzehl, S. Moller, and F. Metze shows the results of implementing a personality evaluation paradigm for spoken input, and it compares human and computer performance in carrying out this task [9]. For this investigation, a professional speaker wrote speeches corresponding to different personality profiles, in accordance with the Five Factor Model questionnaire NEO-FFI. Then, human judges who did not know the speaker estimated the five personality factors. Recordings were also analyzed by using methods based on acoustic and prosodic signals. The results were very consistent between the acted personalities (as evaluated by the judges) and the initial classification of the results. Based on this, the authors concluded that they had made a first step toward the use of personality traits in conversations for future human-machine communication.

The study of A. V. Ivanov, G. Riccardi, *et al.* focused on personality prediction in the context of human spoken conversation [10]. For that purpose, once again, the Five Factor Model was used as a reference. The authors' final goal is to create a machine called the *Personable and Intelligent Virtual Agent*, which is capable of adjusting its linguistic behavior as required by the human with whom it converses. This would facilitate human-machine communication. During

this research work, a simulated tourist help agent was created, which gathered linguistic and acoustic information from the subjects taking part is a role-playing game. These individuals volunteered their scores in the Big Five (Five Factor Model) questionnaire, and they were classified by their traits in a binary fashion: high or low. The results showed that machines can be trained to automatically predict personality traits based on conversations. In addition, statistically significant data were presented for the prediction of traits such as "conscientiousness" and "extroversion."

Linguistic Inquiry and Word Count (LIWC) is private software that analyzes text and calculates the degree to which an individual uses words from different categories, see [11]. A wide variety of sources are used, such as e-mails, transcripts of conversations, speeches, and poems. With LIWC, it is possible to obtain, for example, information about the number of emotionally negative words or self-references used, among many other dimensions of language.

Research on the topic of personality is often focused on one trait in particular: extroversion/introversion. Researchers in this field strive to find personality indicators, with the goal of creating simulated human-machine conversations, instead of focusing their discoveries on the creation of tools for personality profiling and happiness analysis. It is worth mentioning that, with the exception of the works [3], [12] and the LWIC2007 package (2007), all investigations were carried out based on spoken conversations and not on written text, in contrast with this work. In any case, existing research focused on the inference of personality and happiness based on the analysis of written text does not make use of mobile devices as a platform.

Regarding the research works that do focus on the creation of profiling tools, they are all centered on geographic profiling; they do not include personality as a factor in the profiling of the subject. Despite this, these works emphasize the need to combine disciplines to produce their tools. That is the spirit of this project.

III. OBJECTIVES

The main goal of this project is to develop a prototype system that is capable to collect information in written Spanish from different sources of interpersonal communication on a mobile device.

The project consists of a module in which a client application is developed for mobile devices running the Android operating system. This application is in charge of compiling and sending information about the user to a server application, which stores the information as it is received.

Independently of the goals set for this work, and according to advances in joint research with a team of criminologists from the Institute of Forensic Sciences and Security (ICFS), work will begin on a prototype for a classifier module that, by processing the collected data, will search for markers to classify the user according to Eysenck's theory of personality.

For this purpose, a system will be created to classify the user based on previously established principles of analysis and natural language processing.

Why mobile devices?

According to a study carried out by CISCO Systems (2013), in 2016, there will be more mobile devices than people, which means that there will be a large number of potential users for the system. In addition, it is worth mentioning that many of the most commonly used means of communication are concentrated on these devices.

Why Android systems?

There are many reasons to implement this project on Android devices, the first of which is that the Android OS provides programmers with more flexibility for the development of applications because it allows for free access to all device resources: an indispensable requirement for the development of the proposed system.

Additionally, the percentage of mobile devices running Android rose to 84.1% by the middle of 2012, according to a study by the consulting company Kantar, i.e., more than four out of five people in Spain who possess a mobile device have one that runs Android. This allows for wider distribution of the application.

Nevertheless, not all Android devices are useful to us, or at least not all of them can provide us with the same sources of information. Because of this, we will focus on smartphones, the devices through which most interpersonal communication takes place.

Why in Spanish?

For the purpose of analyzing the conduct of an individual through their writings, knowing and being able to analyze the language in which the individual expresses himself or herself is paramount, from a psychological point of view. The mere fact that someone uses certain specific words or expressions gives structure to the subject's personality profile. Because of this, a single language must be selected for the development of the application. For the application to be used by people in other countries, it would need to be adapted to the appropriate socio-linguistic context.

This project is being developed in Spain, so the native language (Spanish) of the potential users has been selected.

IV. THEORETICAL BACKGROUND

The theory of personality by Hans J. Eysenck [1] is based on multidimensional taxonomies of personality. From this point of view, there exist personality traits that allow for the description, and therefore prediction, of human personality and conduct, see [13].

Eysenck recognizes three personality traits: psychoticism, extroversion and neuroticism, giving rise to the acronym in PEN theory. These traits manifest themselves in different types of human behavior:

TABLE I
CHARACTERISTICS THAT DEFINE THE THREE PERSONALITY TRAITS OF THE PEN
MODEL.

Extroversion	Neuroticism	Psychoticism
Sociable	Irrational	Aggressive
Dominant	Inhibited	Cold
Assertive	Taciturn	Egocentric
Active	Emotional	Impersonal
Lively	Tense	Impulsive
Boastful	Anxious	Antisocial
Daring	Depressed	Creative
Carefree	Feeling guilt	Unfeeling
Adventurous	Low self esteem	Harsh

These traits cannot be understood categorically because they are not mutually exclusive. A subject's personality is composed of three independent traits, which must be understood from a dimensional point of view, [13].

Hence, it is important to understand that the three traits are independent, but together, they determine a personality profile corresponding to the idiosyncrasies of the subject. The potential of their combinations cannot be disregarded.

With this model, an underlying biological basis of the three traits is provided. Eysenck believed that the Extroversion-Introversion trait corresponds to cortical arousal. Specifically, it is controlled by the Ascending Reticular Activating System (ARAS). According to the author, extroverts possess a lower degree of cortical arousal, meaning that they present low cortical activation. In contrast, introverts are *a priori* expected to be highly activated. Given the low "internal" activation of extroverts, they would require external and more intense stimulation, whereas introverts are over-activated and do not require external stimulation to maintain a high level of arousal [14].

The Neuroticism-Stability trait is related to the autonomous nervous system, or the limbic system, which is in charge of regulating emotional impulses. Therefore, a highly neurotic individual will have an unstable autonomous nervous system, leading to intense reactions to stimuli. This would explain the variability of mood and anxiety in neurotic subjects. In stable subjects, the exact opposite would be found, [14].

Psychoticism is the most complicated trait within Eysenck's theory, and only recently has some light been shed on its biological nature. Psychoticism has been found to be related to the vulnerability to psychotic disorders, although this does not mean that people with high scores on this trait are certain to suffer from such personality disorders [14]. The Eysenck Personality Questionnaire-Revised (EPQ-R) [4] is currently used to evaluate the traits proposed by Hans. J. Eysenck.

Lastly, it is worth mentioning the relationship of Eyseck's theory with another multi-trait personality model, which is highly favored by the scientific community: the Five Factor Model. This model, also known as "The Big Five" model [8], is based on five fundamental personality traits: Extroversion, Neuroticism, Openness to Experience, Agreeableness and Conscientiousness [13]. These traits are to be evaluated via the

NEO Personality Inventory-Revised (NEO PI-R) [15], or the Big Five Questionnaire (BFQ) [16].

Extroversion and Openness to Experience correspond to the Extroversion trait in PEN theory, Neuroticism has a homologous trait in Eyseck's theory, and Psychoticism would be inversely correlated with Conscientiousness and Agreeableness.

V. TECHNICAL PROPOSAL

In this section, the architecture and design of the system to be developed is presented and the different components of the system are explained.
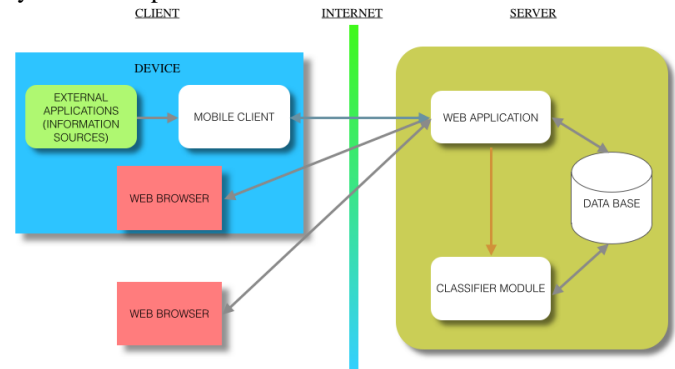

Fig. 1: System architecture

The model to be implemented corresponds to a distributed computer system, which will be composed of numerous devices. Existing classical architectures for distributed systems include the client-server (C/S) architecture and peer-to-peer (P2P) architecture. The C/S architecture is employed when there is a dependency relationship between the devices, which are interconnected in a computer network. This occurs when some functions are performed on the server, and it is the client that communicates with and requests a response from it. In the P2P architecture, every device may function as both client and server.

In the SPD project, there is a logical split within the application. Due to the restrictions described in the non-functional requirements, the system is spread across different computers (physical separation). Only one of the computers— or a group of them functioning as one— will provide services to the rest, thus becoming the "server," the others will submit requests to it, thus becoming "clients." Thus, the chosen architecture is the C/S architecture.

The elements included in the architecture of the SPD system are the following:

- **Client:** software in charge of interacting directly with the user and communicating with the server to submit requests to the system. It will consists of the following:
  - o **Mobile device:** the equipment owned by the user, which contains the following elements:
    - ▪ External applications: an indispensable aspect of the functioning of the system is that the user has a set of applications for interpersonal communication installed on the device, which will serve as the source of information. The

text samples needed for personality profiling will be obtained from these applications. These sources of information will have to be accessible to the mobile client.

▪ Mobile client: application that will be developed in this project, which allows the interaction with the user. It will mainly be in charge of gathering information and communicating with the server to classify the personality traits. It will also work as a client to access the information provided by the external applications.

o **Web client:** some actions will have to be carried out through a web client external to the system. It may be located within the mobile device itself, or on any other device with basic web navigation capabilities.

- **Server:** the software to be implemented will run on a computer that is accessible to the clients installed on various user terminals via the Internet. It will itself be in charge of the task of communicating with the clients, providing services such as registration within the system, setting up user accounts, etc. It contains the following elements, which must be differentiated from the server itself:

o **Web application:** responsible for communicating with both types of clients, mobile and web-based.

o **Database:** element that will store all user access accounts and all information compiled from the mobile client.

o **Classifier module:** module that will take care of processing and determining, from the information stored in the database about a given user, which personality profile defines said user.

In addition, both the client and the server will employ a Model-View-Controller (MVC) architecture. This is an architecture typically used for graphical interfaces such as web pages. It separates the components into different layers for the reuse of code, and its decoupling facilitates the development, expansion and maintenance of the application. In this project, a pure implementation of the model will not be used because it will not always be the user who carries out the actions and requests between clients and server.
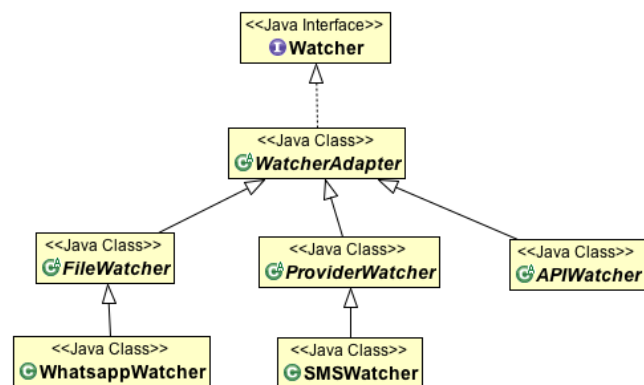

Fig. 2: Details of the hierarchy of implementation of the recollectors.

A. Alternatives

With regard to possible alternatives, we would like to mention that a similar software tool with the same goals as the one presented in this work does not exist, or at least is not publically known. Therefore, this project cannot be implemented using some existing alternative scheme. Notwithstanding, it would be reasonable to carry out an analysis of existing technologies, such as tools or programming languages, to determine which could be useful and to explain why some should be chosen instead of others. We would like to mention that there exists a real-time monitoring software package for Android called *MobileSpy*, which overlaps with SPD as it also collects data. This software, in contrast to the proposed SPD software, collects data from additional sources, such as pictures taken and websites visited by the user. Such information is not necessary for this project, as it is oriented toward espionage.

VI. CLASSIFIER MODULE

In accordance with what was explained in Section III, a prototype classifier module has been programmed as we proceed with the psychological research on the classification and quantification of personality traits through written text.

The classifier is composed of two parts:

First, there is the main process. Its functions include communicating with the database to procure the user data to be processed. In addition, this first part will be in charge of calling the functions of the second part— functions that search for matches of personality indicators— for the evaluation of the established parameters in the compiled user entries. Finally, the main process will have the task of saving in the database the results obtained by the matching functions (for future access, without the need to repeat the analysis on data that have already been processed), and it will be responsible for analyzing the results to evaluate the scores of the different personality traits.

The analysis process is still being developed by the ICFS team and, thus, has not been implemented yet. Basically, they are the search functions that have the task of individually finding and counting every match identified in the data entries, according to the indicators established by the group of psychologists. This group of professionals built a series of lists of indicators to identify personality traits. For example, to identify high scores in the "neuroticism" trait, a list of "emotionally negative words" has been set as an indicator. These indicators are not simply based on the contents of the text, as in the previous example; they also take into account the structure of the text, e.g., lexical density.

Regarding the technical aspects of the classifier, the first decision was to use a language processor to link the analysis to the theoretical guidelines developed by the psychologists. A reference in this field is the *Natural Language Toolkit* (NLTK), but it does not have enough resources to function in the Spanish language. After taking into account many possible solutions and testing a variety of resources to analyze text, we

found that the open source toolkit Freeling was the most appropriate for this project. Freeling offers a wide range of functionalities similar to those of NLTK, but with more resources for Spanish. Furthermore, there exists a translation of the WordNet— a lexical database— called the *Multilingual Central Repository* (MCR), which is compatible with *Freeling* and is on par with numerous resources created by other groups. This helped with the automation of the analysis of text using *Freeling*. Thus, the steps needed to implement the language processor have been determined, and this work may proceed as soon as the psychological part of the research allows for it.

VII. First Prototype: Preliminary results

In this work, the aim is to try to classify messages with within the personality features described in Table 1 using supervised machine learning algorithms. The key idea is that these messages have words that will be preprocessed and clustered in order to see whether its possible to match somehow the obtained clusters with the personality features model described before.

However, although the mobile application described before is already developed and ready to deploy we don't have enough information to analyze nor prove the complete method with Whatsapp and SMS messages. There are not many public datasets available in Spanish, so what we have done as a proof of concept is to work with one public dataset with real messages and two datasets generated from messages collected from different websites. For the public dataset we have used a subset of a corpus made of 63.017 *Twitter* messages released by the end of 2013, 17. This dataset has been analyzed for sentiment/opinion analysis with different techniques by several research institutions (IMDEA, LSI-UNED, Elhuyar Fundazioa, L2F, and SINAI-Universidad de Jaen). Other available dataset in Spanish is the 15M Dataset from Complex Systems and Networks Lab (http://cosnet.bifi.es/research-lines/online-social-systems/15m-dataset). This large dataset of *Twitter* messages from the Spanish 15M movement is not useful for this purpose because it is mainly made of hashtags. The other two datasets are made of messages collected manually from several blogs focusing on collecting different lengths of messages (no rights for publishing). For datasets #1 and #2 a classification for extraversion, neuroticism and psychoticism has been made. Personality (based on Eysenck's theory) could be considered an objective measure, but happiness is usually considered opinion. Therefore, there is not a direct translation between the personality features such as extraversion, neuroticism, and psychoticism into happiness. We know that in PEN model, personality it is measured as a weighted combination of features. If written text can express personality features we think that it can express happiness as well, and for future studies we will try to find if any of these combinations can be somehow correlated into an acceptable happiness classification. For these reasons we have decided to include a happiness classification for dataset #3.

TABLE II
TYPE OF DATASET USED FOR TESTING THE CLUSTERING ALGORITHMS

Dataset	Nº messages	Avg. words / Message	Description
#1	1772	6.53	Twitter
#2	314	10.62	Blogs
#3	200	50.02	Blogs

It is important to remark that the main objective of these experiments is to check if it is possible to identify and classify different types and lengths of messages with common machine learning techniques. In order to do this, we are testing different datasets with several classifying algorithms. First common stage when analyzing texts is to preprocess information. This step will analyze words trying to identify language (English, Spanish, etc.), common lingo, symbols, emoticons, words without semantics and prepositions, articles and conjunctions.

After this initial step, then a stemming analysis takes place. The stemming is a wide used process in information retrieval that reduces or derives words to their stem or root form (if possible). Once this stage finishes, the cleaned dataset is used for clustering.

The first experiments were conducted with WEKA. When working with so many parameters, many different setups can be used. For replication purposes here is a summary of the actions taken, the parameters used and the algorithms tested:

- Preprocessing stage (applied directly to the dataset):
  - Filtering stage (importing data in WEKA):
    - Euclidean distance
    - StringToWordVector filter
    - IDF Transform = True
    - All keywords in lowercase
    - Remove all special characters, punctuation, accents
    - Stopwords: prepositions, articles, conjunctions
    - Eliminate those words that appear less than 5 times.
    - Stemming algorithm: Snowball
    - Deleting keywords with less than 3 letters (before stemming)
    - Deleting numbers
- Used algorithms (WEKA):
  - J48, binary splits, 0.25 confidence factor
  - Random Forest, 30 trees, no depth limit
  - Support Vector Machine (SVM), LibLINEAR

The Table 3 shows the preliminary results with the three different algorithms and random cross-validation (10 folds).

TABLE III
RESULTS FOR ALL SCENARIOS AND ALGORITHMS TESTED

	J48	Random Forest	SVM
#1	Accuracy: 0.42 Kappa: 0.18	Accuracy: 0.57 Kappa: 0.37	Accuracy: 0.56 Kappa: 0.38
#2	Accuracy: 0.65 Kappa: 0.53	Accuracy: 0.58 Kappa: 0.44	Accuracy: 0.65 Kappa: 0.53
#3	Accuracy: 0.64 Kappa: 0.27	Accuracy: 0.64 Kappa: 0.28	Acurracy: 0.82 Kappa: 0.65

The first dataset has in average very short messages for an adequate classification. The accuracy of the tested algorithms in the best case is a 57% (SVM), with many false positives mainly between neuroticism and psychoticism classes (see Table 4 in Appendix). Other issue is the nature of the messages as *Twitter* it is not usually a social media for sending personal messages. Furthermore, most of the selected messages are classified into psychoticism and neuroticism, due to the fact that the dataset comes from primer elections, which bias the type of message collected. These messages are probably really useful for opinion mining but excessively focused on primer elections and short to extract what our psychologists are looking for. The second and third datasets have clearly more promising results. We think that is because messages are longer than the ones from the *Twitter* dataset (see average words per message in Table 2) and because there are not only focused on one specific topic.

VIII. Conclusions

Returning to the goals laid out in Section II, we have created a mobile application for the Android OS that allows for the acquisition of information in the form of written text and allows for communication with a server. In our system the information is acquired from two sources: the *Whatsapp* application and the SMS service. We have also created a server that receives this information from the mobile clients and stores it. Finally, we have also begun work on the creation of a classifier for the gathered information. A prototype has been developed that will continue to be improved.

With regard to the objectives established previously, it can be concluded that all the requirements have been successfully met, except for one: the requirement for the presentation of results in the mobile application. To meet this requirement, it will be necessary to fully develop the classifier and the prototype but this work is still in an early stage although it is a great starting point for future analysis. A lot of work needs to be done prior having more conclusive results. However, so far, we are proud to say that the first objective is achieved. A lot of time has been needed for collaborating with psychologists, understanding the possible ways of identifying personalities and their link with happiness. Finally, developing an application able to collect the *Whatsapp* messages has been a hard task but will allow us the possibility of building the first public corpus of *Whatsapp* messages up to the date. The future of this dataset depends on the overcoming of other problem encountered: the reticence of users to install the application and their non-objective use when installed.

A. Future lines of work

In the short term, we have several goals for continuing the work already done in this project.

First of all, we consider very important to be able to carry out an initial research, so the preliminary results can guide our team and the psychologists to refine the whole system. This first step is already set to be carried out in the near future, the application will be installed on the mobile phones of experimental subjects who will have taken the EPQ-R personality test beforehand. After that, information will be collected and classified by using the tools that have already been developed, which by then will have been improved. The appropriate statistical analysis will be carried out to compare the results of the personality test with the application information. With these results, we will be able to determine how fine-tuned the classifier and the theoretical psychological guidelines are, and we will see what aspects need to be improved so that the personality profile obtained from the application can be trusted. Additionally, we may introduce Artificial Intelligence techniques, such as genetic algorithms, neural networks or automatic learning techniques, as they may help improve the classification.

As mentioned in the previous section, it will be necessary to complete and improve the classifier; this task is intimately related to the research to be carried out. The classifier is a part of the system that cannot be considered finished with only a first build. It will be necessary to verify the proper functioning of the global system and to perform certain adjustments in response to the results. Hence, the system will be improved due to advancements in psychological practice as well as future research that may be carried out.

We are determined to refactor this application and its principles beyond the criminological and clinical fields and to extend the technology to personal use. The field in which we are very interested is human-machine communication. If a machine were able to automatically identify personality from the text provided by a person, it could adapt to the needs and tastes of the person. As [18] mentioned, the moment we are able to identify personality, we are one step closer to being able to simulate it. That is, we could be closer to achieve artificial intelligence.

IX. Acknowledgments

X. Appendix

Detail of confusion matrix obtained with different algorithms and datasets.

TABLE IV
CONFUSION MATRIX FOR TESTS DONE WITH DATASET #1

J48 Confusion Matrix: Dataset #1

Act \ Pred	Extraversion	Neuroticism	Psychoticism	Unknown
Extraversion	**62**	184	48	18
Neuroticism	85	**354**	150	48
Psychoticism	70	138	**247**	115
Unknown	25	19	129	**80**

Random Forest Confusion Matrix: Dataset #1

Act \ Pred	Extraversion	Neuroticism	Psychoticism	Unknown
Extraversion	**114**	187	10	1
Neuroticism	46	**507**	82	2
Psychoticism	1	192	**365**	12
Unknown	1	13	210	**29**

SVM Confusion Matrix: Dataset #1

Act \ Pred	Extraversion	Neuroticism	Psychoticism	Unknown
Extraversion	**199**	102	9	2
Neuroticism	130	**394**	106	7
Psychoticism	22	162	**312**	74
Unknown	1	14	147	**91**

TABLE V
CONFUSION MATRIX FOR TESTS DONE WITH DATASET #2

J48 Confusion Matrix: Dataset #2

Act \ Pred	Extraversion	Neuroticism	Psychoticism	Unknown
Extraversion	**51**	7	22	0
Neuroticism	8	**47**	8	13
Psychoticism	30	12	**30**	7
Unknown	12	14	8	**45**

Random Forest Confusion Matrix: Dataset #2

Act \ Pred	Extraversion	Neuroticism	Psychoticism	Unknown
Extraversion	**52**	6	17	5
Neuroticism	0	**58**	5	13
Psychoticism	32	16	**26**	5
Unknown	9	18	5	**47**

SVM Confusion Matrix: Dataset #2

Act \ Pred	Extraversion	Neuroticism	Psychoticism	Unknown
Extraversion	**49**	3	21	7
Neuroticism	0	**56**	5	15
Psychoticism	16	10	**43**	10
Unknown	2	16	5	**56**

TABLE VI
CONFUSION MATRIX FOR TESTS DONE WITH DATASET #3

J48 Confusion Matrix: Dataset #3

Act \ Pred	Happy	Unhappy
Happy	**670**	330
Unhappy	403	**597**

Random Forest Confusion Matrix: Dataset #3

Act \ Pred	Happy	Unhappy
Happy	**702**	298
Unhappy	417	**583**

SVM Confusion Matrix: Dataset #3

Act \ Pred	Happy	Unhappy
Happy	**844**	156
Unhappy	193	**807**

REFERENCES

[1] H. J. Eysenck, Dimensions of Personality, 1947.
[2] N.D. Livergood, Strategic Personality Simulation: A New Strategic Concept. U.S. Army War College, 1995
[3] Gill, A. and Oberlander, J. [2002] Taking care of the linguistic features of extraversion. In Proceedings of the 24th Annual Conference of the Cognitive Science Society, pp363--368. Washington DC, August 2002
[4] Eysenck, H.J. and Eysenck, S. B. G., 1985. Revised Eysenck Personality Questionnaire (EPQ-R)
[5] G. Young, "Mapping Mayhem: The Geography of Crime," Computoredge, 2003.
[6] K. Rossmo, Profiling and Geography, ISBN-0849381290, 1999.
[7] F. Mairesse, M. and Walker, M., 2006. Automatic Recognition of Personality in Conversation. In: Association for Computational Linguistics ed., Proceedings of the Human Language Technology Conference of the North American New York: Association for Computational Linguistics. Automatic Recognition of Personality in Conversation, pp. 85
[8] Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. In L. Wheeler (Ed.), Review of Personality and Social Psychology, Vol. 2. Beverly Hills, CA: Sage.
[9] T. Polzehl, S. Moller and F. Metze, Automatically Assessing Acoustic Manifestations of Personality in Speech, 2010.
[10] A.V. Ivanov, G. Riccardi, A.J. Sporka and J. Franc, "Recognition of Personality Traits from Human Spoken Conversations," Annual Conference of the International Speech Communication Association, 2011.
[11] J. W. Pennebaker, R.J. Booth and M.E. Francis, Linguistic Inquiry and Word Count (LIWC2007), 2007.
[12] M. de Juan-Espinosa and A. Ortigosa Juárez, Detección Automática De La Personalidad En Comunicación Vía Ordenador. Madrid: Universidad Autónoma de Madrid, n.d.
[13] B. Moreno, Psicología De La Personalidad. Procesos. International Thomson Editores, 2008.
[14] M. de Juan-Espinosa, Personalidad y Criminología, Apuntes De Psicología Criminológica. Madrid: Universidad Autónoma de Madrid, 2005.
[15] Costa, P.T. and McRae, R.R., 1990. Revised NEO Personality Inventory (NEO PI-R).
[16] Caprara, G.V., Barbarranelli, C. and Borgogni, L., 1993. Big Five Questionnaire (BFQ).
[17] Villena-Román, J., García-Morera, J., Moreno-García, C., Ferrer-Ureña, L., Lana-Serrano, S., Carlos González-Cristóbal, J. Westerski, A., Martínez-Cámara, E., García-Cumbreras, M.A., Martín-Valdivia, M.T., Alfonso Ureña-López, L., 2012, TASS-Workshop on Sentiment Analysis at SEPLN, Workshop on Sentiment Analysis, Sociedad Española para el Procesamiento del Lenguaje.

[18] M. de Juan-Espinosa, Personalidad Artificial: Hacia Una Simulación De Las Diferencias De Personalidad En Situaciones De Interacción. Madrid: Universidad Autónoma de Madrid, 1997.

Yago Sáez received the degree in computer engineering in 1999. He got his PhD in Computer Science (Software Engineering) from the Universidad Politécnica de Madrid, Spain, in 2005. He is vice-head of the Computer Science Department from the Carlos III University of Madrid since 2008, where he got his tenure and is associate professor. He belongs to the Evolutionary Computation, Neural Networks and Artificial Intelligence research group (EVANNAI) and takes part of artificial neural network excellence association (REDAF). He also belongs to the Computational Finance and Economics Technical Committee of the IEEE Computational Intelligence Society. He is involved in several competitive national and international research projects in collaboration with public and private institutions. His main research areas encompasses the evolutionary computation, the computational economic and finance applications and the optimization by means of metaheuristics. Since 2012 he has a very active relationship with Stanford University related to auctions.


Carlos Navarro received the degree in computer engineering in 2013 from the Universidad Carlos III de Madrid, Spain. He has worked in IT companies for three years. Since 2012 he has also has been working as a collaborator in the ICFS (Institute of Forensic Science and Security). Now he is studying a Master Degree in Research and Innovation in Information and Communication Technologies and a Master Degree in Computer Engineering in the Universidad Autónoma de Madrid.

Dr. Asuncion Mochon is currently associate professor at the Department of Applied Economics and Economic History at UNED, Madrid, Spain. She received her Ph.D. in Economics with Honors from UNED University in 2003 and her degree in Economics from the Universidad Autonoma de Madrid in 1998. Her research has focused on several topics such as computational economics, auctions and happiness. She has written numerous articles and participated in multiples conferences on these topics. Most of her recent work, both theory and practice, has addressed questions in economics by means of evolutionary computation techniques and game theory analysis.
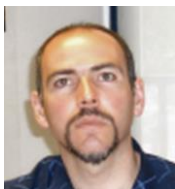
Pedro Isasi: Graduate and Doctor in Computer science by the Polytechnical University of Madrid since 1994. Actually he is University professor and head of the Neural Networks and Evolutionary Computation group of the department of Computer Sciences in the Carlos III of Madrid University. Professor Isasi has been Chair of the Computational Finance and Economics Technical Committee (CFETC) of the IEEE Computational Intelligence Society (CIS). His research is centered in the field of the artificial intelligence, focusing on problems of Classification, Optimization and Machine Learning, fundamentally in Evolutionary Systems, Metaheuristics and artificial neural networks.