

Detecting Narcissist Dark Triad Psychological Traits from Twitter

Lidice Haz^{1,2}^a, Miguel Ángel Rodríguez-García¹^b and Alberto Fernández¹^c

¹Universidad Rey Juan Carlos, Móstoles 28933, Madrid, Spain

²Universidad Estatal Península de Santa Elena, La Libertad, Ecuador

Keywords: Dark Triad, Narcissism, Personality Prediction, Machine Learning, Social Networks, Computational Linguistics, Computational Personality.

Abstract: The fundamental basis of human behavior is personality. This human characteristic influences the tastes and preferences of individuals and the way they interact and communicate with each other. Nowadays, individuals express their feelings, opinions, and ideas to the world by using digital communication platforms such as social media. In recent years, several studies have proposed different models that apply Artificial Intelligence techniques to identify personality traits based mainly on the Big 5 Personality Model and the three subpersonalities of the Dark Triad through the linguistic analysis of their comments online. In this work, we present a study about identifying narcissist dark triad psychological traits. We propose two Machine Learning models to analyze users' behavior in social media. Concretely, we develop a Support Vector Machine and Naïve Bayes method to classify the comments as having non-narcissist or narcissist traits. To train and test the developed method, we have employed NLP techniques to process comments from Twitter and created a manual dataset. Three different techniques have been designed and applied to label each tweet and comment. Then, we conducted several evaluations in which both models reached promising results.


1 INTRODUCTION


Today's society promotes the use of information and communication technologies as fundamental tools for interacting through the web (Fuchs, 2007). Digital communication platforms, and mainly social media, have become prominent means for Internet users to express their ideas, thoughts, opinions and feelings through statuses, comments, and updates (Pratama & Sarno, 2015; Sewwandi et al., 2017; Tadesse et al., 2018).


Physical interaction between people is decreasing as people tend to communicate mainly through virtual media. The anonymity and privacy of these media promote the ability to communicate openly with any user from anywhere and at any time (Baccarella et al., 2018; Sheldon et al., 2019; Van Schaik et al., 2018). This advantage in the use of technology also creates risks. For example, a user can become a victim of sexting, grooming, cyberbullying or scams through dating and relationship websites (Machimbarrena et

al., 2018; Sawyer et al., 2018). In all these scenarios, communication usually begins with a text message through a chat or email. Therefore, it is an arduous task to recognize the real intent of a person. However, it is possible to notice certain parameters of user behavior online by analyzing their written language. Studies in the field of psychology showed that there is a correlation between personality and linguistic behavior of a person (Boyd & Pennebaker, 2017; Pennebaker & King, 1999).

Hence, access to people's public information on social media pages provides important clues about their personality and behavior (H. Ahmad et al., 2020). Understanding user behavior can help to identify personality traits (Adeyemi et al., 2016). Several studies have proposed different techniques to classify online users' personalities by their comments on social media (Hastings et al., 2008; Reidy et al., 2008; Sewwandi et al., 2017). Concretely, in the field of psychology, various personality traits have been described, including the so-called dark triad of

^a <https://orcid.org/0000-0003-1291-1875>

^b <https://orcid.org/0000-0001-6244-6532>

^c <https://orcid.org/0000-0002-8962-6856>

personality: machiavellianism, narcissism, and psychopathy (Jones & Paulhus, 2014). The three constructs are “overlapping, but distinct”. These personalities are used to identify those people who have characteristics or traits that are harmful to society. In other words, the dark triad describes three states belonging to "antisocial" mental schemes (Men, 2014) because they all focus, to varying degrees on social malevolence, on self-promotion, emotional coldness, duplicity, and aggressiveness (Paulhus & Williams, 2002).

In fact, on the one hand, *psychopathy* is marked by high levels of impulsivity and thrill-seeking along with low levels of empathy (Robert D Hare, 1985; Lilienfeld & Andrews, 1996). “Concordant with their impulsive nature, psychopaths constantly seek risky endeavors” (Crysel et al., 2013), thrill (Paulhus & Williams, 2002), and stimulation (R D Hare & Neumann, 2006). Individuals with subclinical tendencies of psychopathy are highly superficial and manipulate others (R D Hare & Neumann, 2006).

On the other hand, *Machiavellianism* describes a manipulative personality trait, whose possessors are cynical, cold, and immoral (Christie & Geis, 2013; Jones & Paulhus, 2009; Rauthmann, 2012). The motives of this behavior are power, money, and status (Paulhus & Williams, 2002). Furthermore, Machiavellians, driven by egotism and competition, show pragmatic, anti-social, manipulative, exploitive, and duplicitous behavior (Jones & Paulhus, 2009; Rauthmann, 2012). They pursue their own goals by cunning deception and opportunism, amoral action, sharp dealing, and hidden agendas (Jones & Paulhus, 2009; Paulhus & Williams, 2002).

Finally, *Narcissistic* personality describes a pervasive pattern of grandiosity, need for admiration, and lack of empathy. These characteristics are denoted in early adulthood and are present in a variety of contexts (Paulhus & Williams, 2002). Narcissism is listed in the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) (Association & others, 2014) which clearly and precisely describes its diagnosed categories. People with narcissist traits tend to view them as intelligent, powerful, physically attractive, unique, and entitled (Bouncken et al., 2020). Clinical characteristics of narcissism include a grandiose sense of self-importance, exaggerated self-esteem, and fantasies of unlimited success and power (Association & others, 2014). Narcissists are disagreeable extraverts, aggressive, and like to debase others (Jonason & Webster, 2012; Paulhus & Williams, 2002). Narcissistic behavior is motivated by self enhancement and personal aspiration power

and admiration (Bouncken et al., 2020; Campbell et al., 2011).

Predicting, classifying, and identifying personality traits from social networks is a trendy research area in Computational Linguistics and Natural Language Processing (NLP). Different machine learning techniques have been applied to identify one of the three subpersonalities of the dark triad in particular psychopathy from user tweets (Moskvichev et al., 2017; Wald et al., 2012). The implementation of automatic systems to classify and predict narcissistic personality traits of the dark triad has been shallowly investigated.

In this sense, in this paper we present a work that applies several machine learning techniques such as Support Vector Machine (SVM) and Naive Bayes as well as NLP methods to classify online comments as a narcissistic or non-narcissistic. The model helps to identify the characteristics of narcissistic language in a text.

The rest of the manuscript is organized as follows: Section 2 analyses relevant works related to identifying the dark triad. Then, in section 3, we present the proposed solution. Here, we first describe the process of creating a dataset for training and testing by collecting comments from Twitter and annotating them. Then, we describe the built pipeline for pre-processing comments and training a binary classifier to automatically identify narcissistic personality traits. Section 4 presents the results of several experiments that we conducted to evaluate the model's precision. Section 5 concludes the paper and analyze future lines of research.

2 RELATED WORKS

In recent years, technological advancement has enabled the development of new ways for analyzing personality. Online spaces have increasingly become a medium for self-expression and social communication. Social media websites allow users to build an online identity, post content (text updates, links or images) and interact with others (Kulkarni et al., 2018). Consequently, new challenges have emerged along with new research lines as Computational Personality Analysis. One of the primary targets of this trend is to automatically identify and classify personal traits by using sophisticated NLP techniques (Celli et al., 2013). Generally, Text Mining, Machine Learning, Information Mining and Computational Linguistics are some of the Artificial Intelligent disciplines used to build the computational prediction model in this topic (Salloum et al., 2017). Several authors have looked at automatic

personality identification through Social Media content coming from Facebook, Twitter and LinkedIn (Alam et al., 2013; de Ven et al., 2017; Lukito et al., 2016). Ahmad & Siddique (2017) used a list of keywords related to the four dimensions of the DiSC model (Dominance, Influence, Submission, Compliance) (Price, 2015). The dataset used was collected from Twitter. The study showed a correlation between vocabulary used and personality type and seated the groundwork for other personality studies using linguistic analysis. Varshney et al. (2017) used The Big Five personality model, which is a set of five broad personality trait dimensions or domains: Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness (Goldberg, 1990). They applied three different classification algorithms namely SVM, K-Nearest Neighbor (KNN) and Multinomial Naïve Bayes (MNB), and they used the concept of "combined results" for identifying the personality traits. Pratama & Sarno (2015) used the myPersonality Dataset and text classification methods (Naive Bayes, KNN and SVM) to identify personality traits from tweets. In the tests, Naive Bayes slightly outperformed the others. The dataset used in this work was obtained from myPersonality project sample data, which offered Facebook users a set of genuine personality and ability measures, and then gives them personalized feedback on their results (Stillwell & Kosinski, 2007). Tandera et al. (2017) did an experiment on personality prediction based on Big Five Personality Model using traditional Machine Learning and Deep Learning to classify the traits. For traditional Machine Learning, they used five algorithms Naive Bayes, SVM, Logistic Regression, Gradient Boosting, and LDA (Differential Language Analysis). In (Golbeck, 2016), a psycholinguistic analysis was carried out calculating the frequency of words according to their meaning as sexual and antisocial words and profanity. They used the tool Linguistic Inquiry and Word Count (LIWC). The study recruited people from Twitter who was previously selected from a psychology site. Then, the Single Item Narcissism Scale (SINS) Test was given to them. As a result, the authors concluded that people who express hateful words or words related to negative emotions and antisocial have a high tendency to narcissism. Therefore, these kinds of expressions are one of the classification parameters for our research. Nevertheless, not all studies utilize social media resources to study personality traits. There are other research areas, such as graphology which utilizes handwriting to analyze these psychology features. In this sense, Wijaya et al., in (Wijaya et al., 2017) proposed a mobile application to predict users'

personalities by analyzing their handwriting. Concretely, the application utilizes an SVM algorithm to classify manuscripts by considering graphology features like page margin. However, those research lines are out of the scope of our work.

In the field of dark triad detection, several studies have been proposed to identify the characteristics of psychopathy (H. Ahmad et al., 2020; Moskvichev et al., 2017). Hancock et al. (2018) propose a study to examine whether people with interpersonally manipulative, callous effect, and criminal tendencies attributes are correlated to specific linguistic patterns. As a result, the authors conclude that linguistic traces of psychopathy can properly be identified in online communication. Wald et al. (2012) applied hybrid techniques for detecting psychopathy from tweets using an ensemble learning technique, known as SelectRUSBoost. The results state that using Select RUS Boost including SVM kernel generated the best result. In other work (H. Ahmad et al., 2020), the researchers implemented a Deep Neural Network model, namely BILSTM for the prediction of psychopathy personality traits regarding online users.

In this context, we observed that The Big Five Personality Model is the most frequently used for identifying personality. Twitter and Facebook are the most common social media sites for extracting data. Manual gathered dataset are commonly the most used and the most common techniques used are machine learning, classification, and linguistic features. However, to the best of our knowledge, neither classification methods nor corpuses have been developed for identifying narcissistic traits in Spanish texts. Therefore, in this work, we propose a model to determine the presence or absence of the narcissistic trait through applying Support Vector Machine and Naïve Bayes methods.

3 THE PROPOSED SOLUTION

This section contains the main development of our proposal. Since there is no availability of a dataset of texts annotated as having or not narcissist traits, we had to create our own dataset for training and evaluating a classifier. Thus, we first present the process followed for generating this data. Then, we describe the classifier.

3.1 Dataset

To drive our study, we created a dataset from Twitter. In this section, we describe the steps we followed to collect and annotate the data.

3.1.1 Triggering Tweets

To collect comments (in Spanish) related to our domain, we published seven contextualized tweets on Twitter. Those seven tweets were related to the seven main traits that characterize the narcissistic personality: authority, self-sufficiency, superiority, exhibitionism, exploitativeness, vanity, and entitlement (R. N. Raskin & Hall, 1979). Besides, before we posted them, two psychologists validated them to ensure that users' reactions to them could unseal the presence narcissistic traits. Table 1 shows three examples of the tweets published (we include their English translation).

Table 1: Example of posted triggering tweets.

Tweet text
<p>¿Cómo crees que actuarías ante el sufrimiento de tu peor enemigo, podrías sentirlo como tuyo, o prefieres cambiar e ignorar el tema? <i>(How do you think you would act in the face of the suffering of your worst enemy, could you feel it as yours, or would you prefer to change and ignore the subject?)</i></p>
<p>¿Escogerías entre la eutanasia de tu madre/ padre por cobrar una herencia? <i>(Would you choose between euthanizing your mother / father to collect an inheritance?)</i></p>
<p>¿Si a un hijo tuyo le detectan una malformación lo darías en adopción, o preferirías que muera para no tener que cuidarlo o que no sufra discriminación? <i>(If your child is found to have a malformation, would you give him/her up for adoption, or would you prefer him/her to die so you do not have to take care of him/her, or he/she does not suffer discrimination?)</i></p>

3.1.2 Announcement

To gain wider visibility and fostering discussion, several messages were spread through posts on Twitter, Facebook, and mailing lists. In those messages, users were asked to participate in the discussion as well as to fill out an NPI test voluntarily, which was available as a web form. We did not collect demographic information or any personal data. We had around four hundred participants recruited in this process that completed the NPI test.

The NPI (Narcissistic Personality Inventory) test is the most widely used instrument for empirical research on narcissism in normal populations. The fundamental objective of the NPI is not to measure narcissism as a personality disorder but to identify the degree to which individuals differ in narcissism as a personality trait (R. N. Raskin & Hall, 1979). Several studies revealed that the NPI is an instrument that has construct validity (Emmons, 1984; García Garduño,

2000; R. Raskin & Terry, 1988; Watson et al., 1984). The version used for this research consists of 40 questions, which allow identify the seven main traits scrutinized before. This result is interpreted as a more significant presence of narcissistic personality traits in an individual. However, it cannot be taken as a diagnosis for NPD (Narcissistic Personality Disorder). Therefore, even someone who scores the highest possible on the NPI does not necessarily have NPD.

3.1.3 Data Collection

We collected comments for about six weeks. During the extraction process, we filter out tweets that were (i) responses, (ii) embedded some media, (iii) had links in the body of the message, (iv) were retweets or (v) has links to newspaper sites.

Since we are analyzing the opinion/reaction of an individual on a particular topic, we decided not to include tweets that follow or publicize non-personal ideas since they are not helpful for our analysis.

3.1.4 Annotation

We applied three evaluations methods to analyze the dataset:

i) We employed the NPI method (Narcissistic Personality Inventory), where a tweet was labelled as narcissistic or non-narcissistic depending on the NPI score obtained by its writer. Therefore, this method could only be applied to tweets created by users who commented any of the seven triggering tweets and fill out the NPI test.

ii) We used a dictionary of words based on a result presented by Golbeck (2016) and the experience of two clinical psychologists. The dictionary contained the most frequent words used by narcissists. For instance, the term "odio" (hate) is frequently used by people with strong narcissistic traits. We utilized such a method due to the limited comments' length, only 240 characters, and the short number of meaningful words such as verbs and qualifying adjectives that could be extracted from comments and could be helpful to unseal some personality traits. Thus, we established the two-word rule, which checked each word of an obtained comment with the words stored in the dictionary. Then, if at least two words were found in the dictionary, the comment was labelled as narcissistic, otherwise non-narcissistic. According to the literature, this rule has been widely applied in several approaches to filter tweets and reduce the number of false positives collected. For instance, Speriosu et al. (2011) consider only two words from a dictionary to filter non-English tweets. Similarly,

Srivastava et al. (2020) utilize such a rule to classify tweets that combine two Hindi languages, Roman and Devanagari. Also, Chellal et al. (2016) use it to select tweets that will be provided to their incremental tweet summarization approach. Conversely, Toh et al. (2015) utilized the rule during the experiments, in the postprocessing phase, to reduce entries in the dataset that do not contain at least two words from various name lists.

iii) Finally, we selected a subset of the comments and drove a manual analysis, which was conducted employing the criteria of two psychologist experts. The process consisted of both experts read, interpreted, and provided an opinion for each comment. Then, if both experts agreed with the opinions of “narcissistic”, the tweet was classified as such. However, if the experts made different opinions, they conducted a new review and discussion until both agreed.

Table 2 details the number of tweets classified as narcissistic and non-narcissistic by each method described above. We have harvested 1092 tweets in total, from which 310 has been labelled as Narcissistic and the remaining as Non-Narcissistic.

Table 2: Dataset Classification Statistic.

Method	Narcissistic	Non-Narcissistic	Total
NPI Test	93	312	405
Dictionary	119	358	477
Manual Evaluation	98	112	210
Total	310	782	1092

3.2 Narcissistic Classifier

The proposed methodology is mainly divided into two steps: (1) pre-processing data, where text data will be represented by utilizing the Vector Space Model; (2) narcissist classifier development, where a supervised model is designed and utilized to automatically identify whether a comment expresses narcissistic traits. Figure 1 shows the system architecture of our detection narcissism model.

As Figure 1 shows, the pipeline of the classifier consists of two main tasks. Firstly, we pre-process the comments to remove punctuation, smileys symbols and stop words. Then, we tokenize the cleaned strings into a set of words that will be filtered by using a Bag of Words (BoW) model. We utilize such a model to extract terms that are used by people who have

narcissism traits, according to two psychologists’ experts. Secondly, we build two different classifiers using machine learning techniques, such as SVM and Naïve Bayes. The filtered word lists feed the algorithms, and they classify if the users’ texts are narcissistic or non-narcissistic.

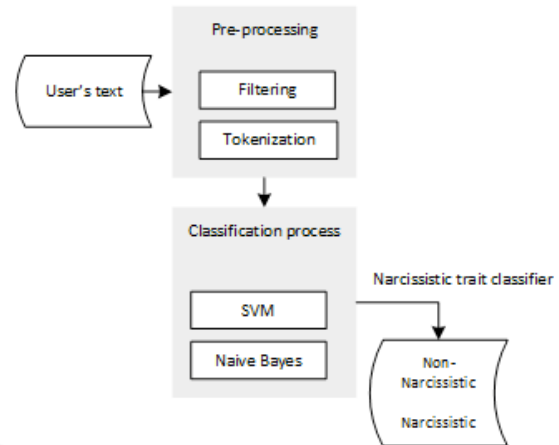


Figure 1: System architecture.

3.2.1 Text Pre-processing

In this phase, we applied some basic pre-processing steps to the acquired data set. People do not always use formal language in their tweets. Thus, we analyze each text and clean it by removing hashtags, mentions, punctuation marks, and accents. Once it is clean, we tokenize it in a set of words. Table 3 shows a comparison between an original and pre-processed tweet. Next, we filter out stop words, i.e. words with no meaning but that are required in the grammatical structure of the language (Leskovec et al., 2020).

Table 3: Comparison of original tweet vs pre-processed tweet.

Original Tweet	Pre-processed Tweet
Es el sufrimiento de mi peor enemigo no podría sentirlo como mío solamente observaría como el karma castiga gente que hace mal .Solamente tendría pena por el o ella 😊	(es, sufrimiento, peor, enemigo, no, podría, sentirlo, observaría, karma, castiga, gente, hace, mal, tendría, pena)

3.2.2 Feature Extraction

In this phase, we use the Bag of Words model to quickly identify words related to a specific domain, such as the most narcissists common words.

The Bag of Words model aims at representing text as occurrences of terms within a document. This representation allows measuring the frequency of each word or having a controlled vocabulary of known words to express sentences by fixed-length vectors. In this work, we have utilized the latter to identify narcissists common words systematically. Hence, we used the aforementioned dictionary built by the experts to identify those words on each tweet. As a result, we obtained a vector that relates the tweet's classification to the number of words inside the pre-defined dictionary. Therefore, when a tweet is analyzed, we tokenize it in a set of words. Then, for each contained word, we analyze whether it appears in the dictionary. If so, the column for this word will be set to one, and zero otherwise. Consequently, we obtain a binary representation in a fixed-length vector of 200, which is the size of the dictionary. Table 4 shows an example of four different tweets in the Bag of Words model. The column *classification* represents how the tweet has been classified as narcissistic or non-narcissistic, and the columns ('ignorar', 'odio', 'soy', 'unico', ...) are the words of the dictionary and the values 1 and 0 indicate whether the word appears in each tweet (rows), respectively.

Table 4: Table of tweets attributes.

classification	ignorar	odio	soy	unico	...
narcissist	1	1	0	0	...
non-narcissist	0	0	0	0	...
narcissist	0	0	1	1	...
non-	0	0	0	0	...
...

3.2.3 Model

The model presented was trained and tested on the Spanish corpus described in section 3.1. We used the machine learning library written in Python to train the model, namely Scikit-Learn (Pedregosa et al., 2011). The algorithms used for the training were: i) Support Vector Machine (SVM), a supervised learning algorithm that analyses the data and recognizes patterns used for classification (Schütze et al., 2008). SVM takes the set of training data and marks it as part of a category, then it predicts whether the test document is a member of an existing class; ii) Naïve Bayes (NB) is a probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions between the features (Duda et al., 1973). Naïve Bayes (NB) is one of the most well-known data mining algorithms for classification (Wu et al., 2008).

To classify the presence or absence of narcissist traits in comments, we utilized a binary classifier. We conducted a series of tests with various scenarios to evaluate the accuracy of the algorithm. For model validation, we used a 5-fold cross-validation technique, applied over all data in each dataset (NPI, Dictionary, Manual and complete).

4 EXPERIMENTS AND RESULTS

In this section, we report and discuss the performances of the developed binary classifier to identify presence or absence of narcissist traits. We have measured the performance of our model using well-known metrics: precision (Prec), recall (Rec), F1-measure (F1) and accuracy (Acc).

The experimental scenarios were performed using the dataset described in section 3.1 as well as its sub-datasets (defined by the annotation method used). The two techniques indicated in section 3.2, namely NB and SVM, were evaluated and compared.

Table 5 presents the results of NB and SVM narcissist classifiers (i.e. Narcissist is the positive class). The results for the complete dataset show that NB outperforms SVM in accuracy (0.79 vs 0.74), precision (0.67 vs 0.54) and F1 (0.56 vs 0.52), while they obtain similar results in recall.

We wanted to analyze whether the annotation method used when constructing the dataset could directly affect the results. Thus, we did experiments using each sub-dataset, which showed the behavior is not very different (SVM only obtains better results in recall and F1 using the manually annotated sub-dataset). Nevertheless, it is noticeable that the low performance of SVM in comparison to NB occur in the NPI-based sub-dataset.

Table 5: Performance measures *Narcissist* classifier (unbalanced dataset).

Narcissist classifier					
Dataset	Classifier	Acc	Prec	Rec	F1
NPI	NB	0.83	0.62	0.70	0.65
	SVM	0.78	0.53	0.39	0.45
Dictionary	NB	0.83	0.70	0.53	0.60
	SVM	0.82	0.70	0.47	0.56
Manual	NB	0.74	0.82	0.56	0.67
	SVM	0.74	0.76	0.64	0.70
ALL	NB	0.79	0.67	0.49	0.56
	SVM	0.74	0.54	0.50	0.52

In Table 6 we present the results when the positive class is *Non-narcissist*. The numbers show again a better performance of NB in most metrics although, in general, the gap between them is narrower. Obviously, the accuracy is the same.

Table 6: Performance measures *Non-narcissist* classifier (unbalanced dataset).

Non-narcissist classifier					
Dataset	Classifier	Acc	Prec	Rec	F1
NPI	NB	0.83	0.91	0.87	0.89
	SVM	0.78	0.84	0.90	0.87
Dictionary	NB	0.83	0.86	0.92	0.89
	SVM	0.82	0.84	0.93	0.88
Manual	NB	0.74	0.69	0.89	0.78
	SVM	0.74	0.72	0.82	0.77
ALL	NB	0.79	0.82	0.91	0.86
	SVM	0.74	0.81	0.83	0.82

The main observation by comparing the narcissist and non-narcissist classifiers is that the latter obtains considerably better results in precision, recall and F1. For example, there is a difference of 0.3 in F1 using the whole dataset (both in NB and SVM).

The reason of this so different performance may be because the dataset used is slightly unbalanced. As shown in Table 2 (section 3.1) there are 310 samples labelled as narcissists and 782 non-narcissists i.e., roughly a 30:70 imbalance ratio. This forces the classifiers to focus on learning how to classify the dominant class (i.e., the class with more examples).

Then, we have carried out experiments with balanced datasets. We applied random under-sampling to find equilibrium on each sub-dataset, i.e. we randomly chose 310 samples of non-narcissist class (93 NPI, 119 Dictionary, 98 manual).

Tables 7 and 8 show the results of the experiments with balanced datasets, for the narcissist and non-narcissist classifiers, respectively. The behavior of the models within their datasets is similar to the unbalanced datasets, with small modifications. That is, in general, NB also outperforms SVM in most metrics. However, the main differences are observed when comparing the results against those obtained with unbalanced datasets. We can observe that now the accuracy has decreased (e.g. 0.73 now vs 0.79 before for NB with whole dataset). However, F1 increases for narcissist classifier (0.72 vs 0.56), while decreasing for non-narcissist (0.72 vs 0.82). These observations confirm the effect of using an

unbalanced dataset, the classifiers learn to identify better the dominant class (non-narcissist).

Nevertheless, these conclusions must be taken with cautiousness due to the limited size of the balanced dataset. Thus, these conclusions need to be revised with an extended dataset.

Table 7: Performance measures *Narcissist* classifier (balanced dataset).

Narcissist classifier					
Dataset	Classifier	Acc	Prec	Rec	F1
NPI	NB	0.85	0.84	0.88	0.86
	SVM	0.79	0.77	0.83	0.80
Dictionary	NB	0.88	0.93	0.83	0.88
	SVM	0.85	0.80	0.93	0.86
Manual	NB	0.76	0.80	0.67	0.73
	SVM	0.78	0.75	0.84	0.79
ALL	NB	0.73	0.73	0.71	0.72
	SVM	0.70	0.69	0.74	0.72

Table 8: Performance measures *Non-narcissist* classifier (balanced dataset).

Non-narcissist classifier					
Dataset	Classifier	Acc	Prec	Rec	F1
NPI	NB	0.85	0.86	0.83	0.84
	SVM	0.79	0.81	0.74	0.77
Dictionary	NB	0.88	0.85	0.93	0.89
	SVM	0.85	0.92	0.77	0.84
Manual	NB	0.76	0.72	0.84	0.78
	SVM	0.70	0.81	0.71	0.76
ALL	NB	0.73	0.73	0.74	0.73
	SVM	0.70	0.72	0.66	0.69

To improve the results in all datasets it is convenient to implement other NLP techniques that allow the textual analysis of the comment and the frequency of words. It is also possible to include the emoticons denoting positive or negative emotions contained in the comments.

The present study showed that it is possible identifying effectively narcissistic traits focusing on specific linguistic features. The work presented by Sumner et al. (2012) shows that there is a high correlation between the words that users use on Twitter with the personality they have. Also, several

works cited in this research have identified psychological traits applying machine learning different techniques such as SVM, KNN, Naive Bayes, Logistic Regression, CNN. The results obtained by these researchers showed an accuracy of about 0.75, which suggests positive results for our research, mainly using Spanish text for identifying narcissistic traits which, as far as we know, no exploration has been conducted. Other approaches correspond to the identification of personality traits (Tandera et al., 2017) and psychopathy dark traits (Wald et al., 2012). To the best of our knowledge, neither classification methods nor corpus have been developed for identifying narcissistic traits in natural text. This study is an initial attempt towards 'people profiling with traits narcissist' with the help of tag words that can be a useful tool in many other areas such as marketing, promotions, advertising, sales, IT applications, anthropological studies, and social media, etc.

5 CONCLUSIONS

Personality traits can be predicted by user-generated content on social media. As we studied in the literature review, personality traits analysis is attracting interest due to the exponential use of social media (Arnoux et al., 2017; Pratama et al., 2015). Twitter is a powerful source of information about a person's psychological individuality. Identifying the personality trait of the dark triad narcissistic from online content is a challenging problem.

This work captured and analyzed tweets with a linguistic perspective. The research was based on background studies on identifying narcissistic traits (García Garduño, 2000; Golbeck, 2016; R. Raskin & Terry, 1988) and the validation of two clinical psychologists. The objective was to train an automatic classification model for identifying the narcissist traits' presence or absence using supervised machine learning algorithms such as the SVM and Naive Bayes. We created a dataset by gathering comments (in Spanish) from Twitter to experiment with the proposed method. We applied different (semi-automatic and manual) techniques to label the collected messages, namely an NPI test, a dictionary of narcissistic words and a manual review.

The results of our experiments showed a promising approach for predicting the traits of the dark triad narcissistic. The highest accuracy (0.8) obtained was using the Naive Bayes and dataset validated with the dictionary. The lower results were harvested using the mixed dataset.

We concluded that Naïve Bayes outperforms SVM in most metrics and experiments, obtaining an accuracy of 79% using the complete dataset for identifying narcissist traits in short texts. While this value cannot be compared to other works with the same goal (we did not find other systems), this performance result is in line to other works that apply machine learning to identify other psychological traits. In this research, we presented a fundamental idea, but there is some work in progress, from which we will try to increase the performance of the classifiers presented.

For future work, the study will consider increasing the scale of the dataset by adding new samples. Also, improve the classifier to identify not only whether the texts contain narcissist traits, but also to recognize which of the seven traits of the narcissistic dark triad is present. In addition, we are considering improving the preprocessing of texts. For instance, we plan to employ semantic approaches to analyze a word not as an isolated entity but in its context. We plan to analyze the performance of other supervised and non-supervised learning algorithms in the problem facing in this work.

ACKNOWLEDGEMENTS

Work partially supported by the Santa Elena Peninsula State University (UPSE), project research "DTC" [91870000.0000.386468], the Community of Madrid, through the Young Researchers R+D Project. Ref. M2173 – SGTRS (co-funded by Rey Juan Carlos University); and by the Spanish Ministry of Science, Innovation and Universities, through grant RTI2018-095390-B-C33(MCI/AEI/FEDER, UE).

The authors would like to thank the collaboration of the psychologists Ivette Gómez and Sara Yagual.

REFERENCES

- Adeyemi, I. R., Abd Razak, S., & Salleh, M. (2016). *Understanding online behavior: exploring the probability of online personality trait using supervised machine-learning approach*. *Frontiers in ICT*, 3, 8.
- Ahmad, H., Arif, A., Khattak, A. M., Habib, A., Asghar, M. Z., & Shah, B. (2020). Applying deep neural networks for predicting dark triad personality trait of online users. *2020 International Conference on Information Networking (ICOIN)*, 102–105.
- Ahmad, N., & Siddique, J. (2017). Personality assessment using Twitter tweets. *Procedia Computer Science*, 112, 1964–1973.

- Alam, F., Stepanov, E. A., & Riccardi, G. (2013). Personality traits recognition on social network-facebook. *Seventh International AAAI Conference on Weblogs and Social Media*.
- Arnoux, P. H., Xu, A., Boyette, N., Mahmud, J., Akkiraju, R., & Sinha, V. (2017). 25 tweets to know you: A new model to predict personality with social media. *In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 11, No. 1)*.
- Association, A. P., & others. (2014). *Guía de consulta de los criterios diagnósticos del DSM-5®: Spanish Edition of the Desk Reference to the Diagnostic Criteria From DSM-5®*. American Psychiatric Pub.
- Baccarella, C. V., Wagner, T. F., Kietzmann, J. H., & McCarthy, I. P. (2018). Social media? It's serious! Understanding the dark side of social media. *European Management Journal*, 36(4), 431–438.
- Bouncken, R., Cesinger, B., & Tiberius, V. (2020). Narcissism, Machiavellianism, and psychopathy of top managers: Can Entrepreneurial Orientation secure performance? *International Journal of Entrepreneurial Venturing*, 12(3), 273–302.
- Bermúdez, J., Pérez, A. M., & Sanjuán, P. (2017). *Psicología de la personalidad: Teoría e investigación. volumen I*. Editorial UNED.
- Boyd, R. L., & Pennebaker, J. W. (2017). Language-based personality: a new approach to personality in a digital world. *Current Opinion in Behavioral Sciences*, 18, 63–68.
- Campbell, W. K., Hoffman, B. J., Campbell, S. M., & Marchisio, G. (2011). Narcissism in organizational contexts. *Human Resource Management Review*, 21(4), 268–284.
- Celli, F., Pianesi, F., Stillwell, D., & Kosinski, M. (2013). Workshop on computational personality recognition: Shared task. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1).
- Chellal, A., Boughanem, M., & Dousset, B. (2016). Multi-criterion Real Time Tweet Summarization Based upon Adaptive Threshold. *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pp. 264–271.
- Christie, R., & Geis, F. L. (2013). *Studies in machiavellianism*. Academic Press.
- Crysel, L. C., Crosier, B. S., & Webster, G. D. (2013). The Dark Triad and risk behavior. *Personality and Individual Differences*, 54(1), 35–40.
- De Ven, N., Bogaert, A., Serlie, A., Brandt, M. J., & Denissen, J. J. A. (2017). Personality perception based on LinkedIn profiles. *Journal of Managerial Psychology*.
- Duda, R. O., Hart, P. E., & Others. (1973). Pattern classification and scene analysis. *Wiley New York*.
- Emmons, R. A. (1984). Factor analysis and construct validity of the narcissistic personality inventory. *Journal of Personality Assessment*, 48(3), 291–300.
- Fuchs, C. (2007). *Internet and society: Social theory in the information age*. Routledge.
- García Garduño, J. M. (2000). La medición empírica del narcisismo: una síntesis de la investigación sobre su relación con rasgos y teorías de la personalidad. *Psicol. Conduct*, 33–56.
- Golbeck, J. (2016). Negativity and anti-social attention seeking among narcissists on Twitter: A linguistic analysis. *First Monday*.
- Goldberg, L. R. (1990). An alternative" description of personality": the big-five factor structure. *Journal of Personality and Social Psychology*, 59(6), 1216.
- Hancock, J. T., Woodworth, M., & Boochever, R. (2018). Psychopaths online: The linguistic traces of psychopathy in email, text messaging and Facebook. *Media and Communication*, 6(3), 83–92.
- Hare, R. D., & Neumann, C. S. (2006). *The PCL-R assessment of psychopathy. Development, structural properties, and new directions [w:] CJ Patrick (ed.), Handbook of psychopathy*, 58–88. New York: The Guilford Press.
- Hare, Robert D. (1985). Comparison of procedures for the assessment of psychopathy. *Journal of Consulting and Clinical Psychology*, 53(1), 7.
- Hassanein, M., Hussein, W., Rady, S., & Gharib, T. F. (2018). Predicting personality traits from social media using text semantics. *2018 13th International Conference on Computer Engineering and Systems (ICCES)*, 184–189.
- Hastings, M. E., Tangney, J. P., & Stuewig, J. (2008). Psychopathy and identification of facial expressions of emotion. *Personality and Individual Differences*, 44(7), 1474–1483.
- Jonason, P. K., & Webster, G. D. (2012). A protean approach to social influence: Dark Triad personalities and social influence tactics. *Personality and Individual Differences*, 52(4), 521–526.
- Jones, D. N., & Paulhus, D. L. (2009). *Machiavellianism*.
- Jones, D. N., & Paulhus, D. L. (2014). Introducing the short dark triad (SD3) a brief measure of dark personality traits. *Assessment*, 21(1), 28–41.
- Kulkarni, V., Kern, M. L., Stillwell, D., Kosinski, M., Matz, S., Ungar, L., Skiena, S., & Schwartz, H. A. (2018). Latent human traits in the language of social media: An open-vocabulary approach. *PloS One*, 13(11), e0201703.
- Leskovec, J., Rajaraman, A., & Ullman, J. D. (2020). *Mining of massive data sets*. Cambridge university press.
- Lilienfeld, S. O., & Andrews, B. P. (1996). Development and preliminary validation of a self-report measure of psychopathic personality traits in noncriminal population. *Journal of Personality Assessment*, 66(3), 488–524.
- Lukito, L. C., Erwin, A., Purnama, J., & Danoekoesoemo, W. (2016). Social media user personality classification using computational linguistic. *2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 1–6.
- Machimbarrena, J. M., Calvete, E., Fernández-González, L., Álvarez-Bardón, A., Álvarez-Fernández, L., & González-Cabrera, J. (2018). Internet risks: An overview of victimization in cyberbullying, cyber dating abuse, sexting, online grooming and problematic

- internet use. *International Journal of Environmental Research and Public Health*, 15(11), 2471.
- Men, I. (2014). *Understanding the Dark Triad—A General Overview*.
- Moskvichev, A., Dubova, M., Menshov, S., & Filchenkov, A. (2017). Using linguistic activity in social networks to predict and interpret dark psychological traits. *Conference on Artificial Intelligence and Natural Language*, 16–26.
- Paulhus, D. L., & Williams, K. M. (2002). The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality*, 36(6), 556–563.
- Predregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & others. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12, 2825–2830.
- Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6), 1296.
- Pratama, B. Y., & Sarno, R. (2015). Personality classification based on Twitter text using Naive Bayes, KNN and SVM. *2015 International Conference on Data and Software Engineering (ICoDSE)*, 170–174.
- Price, L. A. (2015). Disc instrument validation study. In *Texas*.
- Raskin, R. N., & Hall, C. S. (1979). A narcissistic personality inventory. *Psychological Reports*.
- Raskin, R., & Terry, H. (1988). A principal-components analysis of the Narcissistic Personality Inventory and further evidence of its construct validity. *Journal of Personality and Social Psychology*, 54(5), 890–902.
- Rauthmann, J. F. (2012). The Dark Triad and interpersonal perception: Similarities and differences in the social consequences of narcissism, Machiavellianism, and psychopathy. *Social Psychological and Personality Science*, 3(4), 487–496.
- Reidy, D. E., Zeichner, A., Hunnicutt-Ferguson, K., & Lilienfeld, S. O. (2008). Psychopathy traits and the processing of emotion words: Results of a lexical decision task. *Cognition and Emotion*, 22(6), 1174–1186.
- Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2017). A survey of text mining in social media: facebook and twitter perspectives. *Adv. Sci. Technol. Eng. Syst. J.*, 2(1), 127–133.
- Sawyer, A. N., Smith, E. R., & Benotsch, E. G. (2018). Dating application use and sexual risk behavior among young adults. *Sexuality Research and Social Policy*, 15(2), 183–191.
- Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval* (Vol. 39). Cambridge University Press Cambridge.
- Sewwandi, D., Perera, K., Sandaruwan, S., Lakchani, O., Nugaliyadde, A., & Thelijjagoda, S. (2017). Linguistic features based personality recognition using social media data. *2017 6th National Conference on Technology and Management (NCTM)*, 63–68.
- Sheldon, P., Rauschnabel, P., & Honeycutt, J. M. (2019). The dark side of social media: Psychological, managerial, and societal perspectives. Academic Press.
- Speriosu, M., Sudan, N., Upadhyay, S., & Baldrige, J. (2011). Twitter polarity classification with label propagation over lexical links and the follower graph. *First Workshop on Unsupervised Learning in NLP (Conference on Empirical Methods in Natural Language Processing, EMNLP '11)*, 53–63, Edinburgh, Scotland, UK, 53–63.
- Srivastava, A., Bali, K., & Choudhury, M. (2020). Understanding Script-Mixing: A Case Study of Hindi-English Bilingual Twitter Users. 4th Workshop on Computational Approaches to Code Switching, 36–44.
- Stillwell, D. J., & Kosinski, M. (2007). mypersonality research wiki. <https://www.psychometrics.cam.ac.uk/productservices/mypersonality>
- Sumner, C., Byers, A., Boochever, R., & Park, G. J. (2012). Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In *2012 11th international conference on machine learning and applications (Vol. 2, pp. 386-393)*. IEEE.
- Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2018). Personality predictions based on user behavior on the facebook social media platform. *IEEE Access*, 6, 61959–61969.
- Tandera, T., Suhartono, D., Wongso, R., Prasetyo, Y. L., & others. (2017). Personality prediction system from facebook users. *Procedia Computer Science*, 116, 604–611.
- Toh, Z., Chen, B., & Su, J. (2015, July). Improving twitter named entity recognition using word representations. In *Proceedings of the Workshop on Noisy User-generated Text (pp. 141-145)*.
- Van Schaik, P., Jansen, J., Onibokun, J., Camp, J., & Kusev, P. (2018). Security and privacy in online social networking: Risk perceptions and precautionary behaviour. *Computers in Human Behavior*, 78, 283–297.
- Varshney, V., Varshney, A., Ahmad, T., & Khan, A. M. (2017). Recognising personality traits using social media. *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, 2876–2881.
- Wald, R., Khoshgoftaar, T. M., Napolitano, A., & Sumner, C. (2012). Using Twitter content to predict psychopathy. *2012 11th International Conference on Machine Learning and Applications*, 2, 394–401.
- Watson, P. J., Grisham, S. O., Trotter, M. V., & Biderman, M. D. (1984). Narcissism and empathy: Validity evidence for the Narcissistic Personality Inventory. *Journal of Personality Assessment*, 48(3), 301–305.
- Wijaya, W., Tolle, H., & Utaminigrum, F. (2017). Personality analysis through handwriting detection using android based mobile device. *Journal of Information Technology and Computer Science*, 2(2).
- Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., & Others. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1), 1–37.