Predicting Temperament using Keirsey's Model for Portuguese Twitter Data

Cristina Fátima Claro, Ana Carolina E. S. Lima and Leandro N. de Castro
Institute of Problem Solving, Presbiteriana Mackenzie University, Rua da Consolação-930, São Paulo, Brazil

Keywords: Machine Learning, Social Media, Temperament's Classification, Keirsey Temperament Model.

Abstract: Temperament is a set of innate tendencies of the mind related with the processes of perceiving, analyzing and decision making. The purpose of this paper is to predict the user's temperament based on Portuguese tweets and following Keirsey's model, which classifies the temperament into artisan, guardian, idealist and rational. The proposed methodology uses a Portuguese version of LIWC, which is a dictionary of words, to analyze the context of words, and supervised learning using the KNN, SVM and Random Forest algorithms for training the classifiers. The resultant average accuracy obtained was 88.37% for the artisan temperament, 86.92% for the guardian, 55.61% for the idealist, and 69.09% for the rational. By using binary classifiers the average accuracy was 90.93% for the artisan temperament, 88.98% for the guardian, 51.98% for the idealist and 71.42% for the Rational.

1 INTRODUCTION

A set of characteristics is defined according with the personality and these describe the individual behavior, the temperament and the emotion (Nor Rahayu et al., 2016). Personality represents the mixture of characteristics and qualities that builds the character of an individual. Thus, personality prediction is of interest in the areas of health, psychology, human resources and also has many commercial applications. Several researches investigate the link between human behavior in social media, personality types and psychological illnesses, such as depression and post-traumatic stress (Plank and Dirk, 2015; Lima and de Castro, 2016).

Social media are composed of different types of social sites, including traditional media, such as newspaper, radio and television, as well as non-traditional media, such as Facebook, Twitter and others (Gundecha and Liu, 2012). Social media mining is the process that allows the analysis and extraction of patterns from social media data (Nor Rahayu et al., 2016). In this context, this paper develops a system to predict the temperament of Twitter users, using tweets in the Portuguese language. The temperament model used was introduced by David Keirsey, and divides the temperament into four categories: artisan; guardian; idealist; and rational. In order to do so, we will use the TECLA framework adapted to work with Portuguese texts (Lima and de Castro, 2016). In addition, it will be shown an analysis of the context of words by temperament using the dictionary of words Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015).

This paper is structured as follows. Section II presents the David Keirsey temperament model used in TECLA, and Section III describes the TECLA framework. Section IV presents the methodology and the results achieved and, finally, Section V concludes and discusses future perspectives.

2 KEIRSEY'S TEMPERAMENT MODEL

Temperament is a set of innate tendencies of the mind that relates to the processes of perceiving, analyzing, and decision making (Calegari and Gemignani, 2006). People seek success, happiness, love, pleasure, etc., in different ways and with distinct intensities and, therefore, there are different types of temperament (Hall et al., 2000).

The temperament has its history marked in the proposal of the four humors described by Hippocrates, which gave origin to the theory of the four humors to interpret the state of health and illness of a person (Hall et al., 2000). From this theory, Galen (190 AD)

Claro, C., Lima, A. and Castro, L.
Predicting Temperament using Keirsey's Model for Portuguese Twitter Data.
DOI: 10.5220/0006700102500256
In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018) - Volume 2*, pages 250-256
ISBN: 978-989-758-275-2

created the model of the first temperament typology (Ito and Guzzo, 2002).

David Keirsey, an American psychologist, directed his studies to temperament in action, paying attention to choices, behavior patterns, congruencies, and consistencies. For Keirsey, psychological types are driven by aspirations and interests, which motivate us to live, act, move, and play a role in society (Lima and de Castro, 2016; Keirsey, 1998; Calegari and Gemignani, 2006).

The artisans are usually impulsive, they speak what comes to their minds and tend to do what works; whereas the guardians speak mainly of their duties and responsibilities, and how well they obey the laws. Idealists normally act from a good conscience and the rationals are pragmatic, act efficiently to reach their objectives, sometimes ignoring the rules and conventions if necessary (Keirsey, 1996; Lima, 2016).

Keirsey's temperament can be obtained by mapping the result of the MBTI test (Myers-Briggs Type Indicator), which uses four dimensions to classify users, totaling 16 psychological types (Keirsey, 1998; Calegari and Gemignani, 2006; Plank and Dirk, 2015). Psychological types are acronyms formed by letters that begin with E and I, (extraversion and introversion), which are attitudes; S and N indicate sensation and intuition, which is the process of perception; the letters T and F indicate thinking and feeling, and usually use logical reasoning, think first and feel later; and letters J and P indicate judgment and perception, which are attitudes and reflect the individuals' style in the external world (Hall et al., 2000).

The mapping of the MBTI into the Keirsey's model occurs by means of the classification of the acronyms defined by Myers-Briggs, as shown in Table 1 (Keirsey, 1998).

Table 1: Keirsey temperament model classification from the MBTI.

Keirsey	**Myers-Briggs**			
Artisan	ESTP	ISTP	ESFP	ISFP
Guardian	ESTJ	ISTJ	ESFJ	ISFJ
Idealist	ENFJ	INFJ	ENFP	INFP
Rational	ENTJ	INTJ	ENTP	INTP

3 THE TECLA FRAMEWORK

The TECLA framework (Temperament Classification Framework) was developed by Lima & de Castro (Lima, 2016; Lima and de Castro, 2016) with the objective of offering a modular tool for the classification of temperaments based on the Keirsey and Myers-Briggs models (Lima, 2016). It is structured in a modular form, giving greater independence for each stage of the process and making it possible to couple and test different techniques in each module (Lima, 2016). Figure 1 shows the TECLA's modules, which are detailed in the following.

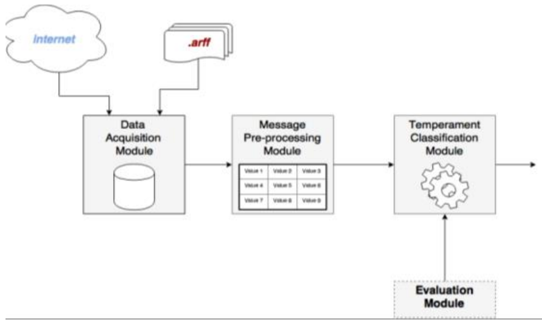

Figure 1: The TECLA framework structure.

- Data Acquisition Module: Receives information from the user to be classified, including the number of tweets, the number of followers and followed, and a set of messages (tweets) from the user;
- Message Pre-Preprocessing Module: Processes the data by creating an object matrix (meta-base) represented by meta-attributes. The information in the TECLA are divided into two categories: grammatical and behavioral. The behavior category uses information from Twitter, such as number of tweets, number of followed, followers, favorites, and number of times the user has been favorited. The grammar category uses information from LIWC, MRC, Taggers, or oNLP (Lima and de Castro, 2016);
- Temperament Classification Module: Responsible for identifying the temperament of social media users. It performs the classification in the Keirsey model by using a set of classifiers;
- Evaluation Module: Used to quantify the framework performance (Lima and de Castro, 2016).

In the version proposed in this paper, the TECLA will be adapted to work with texts written in Portuguese and will use the information provided by the LIWC (Pennebaker et al., 2015).

4 METODOLOGY AND RESULTS

The description to be presented in this section will follow the modular structure of the TECLA framework. First we will explain how we implemented each module of the framework and then the computational results.

A. DATA ACQUISITION

To validate this work we used a data from the literature called Twisty, which has tweets in Portuguese and is provided by CLiPS (The Computational Linguists & Psycholinguistics Research Center) (Verhoeven et al., 2016). The dataset is composed of: user id; tweet id; other tweets id; confirmed tweets id; Myers-Briggs Type Indicators (MBTI) result; and gender. The tweets were captured by using the Twitter API (Xavier and Carvalho, 2011), and we captured the tweets, number of followers, number of favorites, total number of tweets and total number of friends (Kwak et al., 2010) of each user.

The original database consists of 4,090 user ids. From this universe it was not possible to collect 222 user ids due to denied access, leaving 3,868 valid user ids. Table 2 shows the descriptive analysis of the database according to David Keirsey's model, where Tweets_Statuses_Count refers to the number of tweets from the opening of the user account, and Tweets_base refers to the number of tweets collected.

Figure 2 shows the temperament distribution of the users. It is noted that the idealist temperament is the predominant one, totaling 44% of the database.

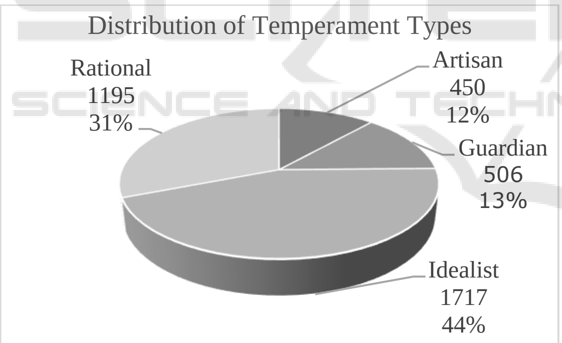

Figure 2: Distribution of the users by temperament.

B. PRE-PROCESSING AND CATEGORY ANALYSIS

At this stage the texts are prepared for the application of the classification algorithms, which consists of the removal of special characters, blank spaces, numbers, symbols, URLs, tokenization, and stopwords removal (Haddi et al., 2013; Spencer and Uchyigit, 2012). After that, a bag-of-words technique was applied to specify the importance of each attribute (token) by assigning a weight to each token based on its TF-IDF (Feldman and Sanger, 2007).

Another way to structure documents is through the use of dictionaries, such as the Linguistic Inquiry and Word Count (LIWC), which allows the grouping of words into psychologically meaningful categories. The LIWC was created by Dr. James Pennebaker to examine relationships between language and personality (Komisin and Guinn, 2012; Pennebaker and King, 1999). It is a textual analysis tool that structures documents into categories by assigning each word to the corresponding category (Pennebaker et al., 2015).

By using the Portuguese LIWC dictionary it was calculated the frequency of words by temperament. The goal here is to present the most spoken word categories for each temperament, as shown in Table 3. In a first analysis it is observed that the rational temperament usually has a higher average frequency of categories, followed by the guardian temperament.

In a second analysis it is noticed that all the temperaments have a higher frequency in the following categories: funct, which are functional words like for, not, very and others; pronoun, which are the pronouns like I, mine, me and others; verb, which are verbs such as cover, occur and other; social, which are social processes such as talking, accompanying and other; cogmech, which is the cognitive category; and relativ, which is relativity as are (Pennebaker et al., 2015) a, turn, exit and other. There are users who tend to write by hiding their identity and tend to present a writing that expresses action and can show a greater perception and logical reasoning.

In another analysis, it is noted prominence for the ppron category for rational and idealist temperaments; and ipron, present, preps, incl for guardian and rational temperaments. These categories are related to the linguistic dimensions that tend to write more pa-

Table 2: Temperament and Twitter data of the users (A = artisan, G = guardian, I = idealist, R = rational).

	A	G	I	R	Total
Users	**450**	**506**	**1.717**	**1.195**	**3.868**
Tweets_Statuses_Count	12.343.807	15.648.860	65.593.286	45.198.150	**138.784.103**
Tweets_Base	674.211	738.755	2.570.646	1.751.624	**5.735.236**
Followers	292.413	423.549	1.497.093	1.799.686	**4.012.741**
Friends	168.893	225.371	825.969	640.529	**1.860.762**
Favorites	1.768.903	2.371.924	10.006.749	6.683.984	**20.831.560**

pers, pronouns, auxiliary verbs, etc. The humans category has higher frequency for the idealist and rational temperaments, and tends to write about people. The affect category occurs more for the guardian and idealist temperaments, and users tend to write presenting positive, negative, and other emotions. Last, the ingest category has higher average frequency for the artisan and guardian temperaments, who tend to write about body, health and others.

With a lower frequency of words and with the same value in all the temperaments the future, family, anx, see, health, death, assent and filler categories stand out.

Table 3: Average frequency of LIWC categories by temperament.

Category	Artisan	Guardian	Idealist	Rational
funct	5,22	5,31	5,30	5,39
pronoun	1,85	1,86	1,91	1,93
ppron	1,21	1,22	1,25	1,26
i	0,39	0,36	0,42	0,41
we	0,04	0,04	0,04	0,03
you	0,66	0,70	0,66	0,68
shehe	0,64	0,67	0,65	0,67
they	0,12	0,13	0,13	0,13
ipron	1,22	1,25	1,24	1,27
article	0,77	0,80	0,78	0,81
verb	1,87	1,88	1,88	1,91
auxverb	0,67	0,69	0,68	0,70
past	0,42	0,43	0,43	0,43
present	1,12	1,13	1,12	1,15
future	0,08	0,08	0,08	0,08
adverb	0,48	0,50	0,50	0,50
preps	1,48	1,52	1,46	1,49
conj	0,91	0,90	0,92	0,93
negate	0,24	0,25	0,25	0,25
quant	0,59	0,60	0,61	0,62
number	0,14	0,14	0,15	0,15
swear	0,71	0,71	0,72	0,73
social	2,16	2,17	2,21	2,24
family	0,04	0,04	0,04	0,04
friend	0,11	0,10	0,10	0,09
humans	1,11	1,11	1,15	1,14
Affect	1,08	1,10	1,09	1,08
Posemo	0,70	0,72	0,71	0,69
negemo	0,35	0,34	0,36	0,36
anx	0,05	0,05	0,05	0,05
anger	0,14	0,13	0,14	0,15
sad	0,17	0,16	0,17	0,17
cogmech	4,06	4,14	4,12	4,18
insight	0,72	0,73	0,74	0,75
cause	0,48	0,50	0,49	0,50
discrep	0,68	0,68	0,69	0,70
tentat	0,98	0,99	1,00	1,02
certain	0,38	0,39	0,39	0,39
inhib	0,53	0,55	0,54	0,55
incl	1,40	1,42	1,40	1,41
excl	0,79	0,80	0,80	0,82
percept	0,78	0,79	0,80	0,79
see	0,26	0,26	0,26	0,26
hear	0,18	0,17	0,18	0,18
feel	0,31	0,32	0,31	0,31
bio	0,71	0,69	0,71	0,72
body	0,32	0,31	0,31	0,32
health	0,13	0,13	0,13	0,13
sexual	0,21	0,20	0,21	0,21
ingest	1,08	1,10	1,06	1,06
relativ	2,30	2,36	2,28	2,31
motion	0,76	0,76	0,74	0,76
space	0,98	1,01	0,97	0,99
time	0,98	1,02	0,96	0,96
work	0,23	0,25	0,23	0,24
achieve	0,46	0,49	0,46	0,47
leisure	0,31	0,32	0,31	0,31
home	0,06	0,06	0,06	0,05
money	0,29	0,30	0,29	0,31
relig	0,08	0,09	0,09	0,08
death	0,06	0,06	0,06	0,06
assent	0,13	0,13	0,13	0,13
nonfl	0,26	0,27	0,26	0,27
filler	0,03	0,03	0,03	0,03

C. CLASSIFICATION

In the experiments carried out, 4% of the total dataset was randomly sampled to be used due to the size of the matrix to be processed and the unavailability of computational resources. To perform the temperament classification, the following classifiers available in the Scikit-learn (Pedregosa et al., 2011) were used: KNN; SVM; and Random Forest (Lima and de Castro, 2016; Nor Rahayu et al., 2016). Each temperament was divided into a binary problem, as proposed by Lima and de Castro (Lima and de Castro, 2016). For the tests, a cross-validation with 6 and 10-folders was used, and the accuracy, precision, recall, and F-measure were calculated. For the KNN classifier, which uses the object classification according to the K-nearest neighbors, K = 1, K = 2 e K = 3 and the cosine similarity was used for determining the neighbors. The tests were separated into LIWC word dictionary LIWC (Pennebaker et al., 2015) and TF-IDF (Feldman and Sanger, 2007).

1) LIWC

Table 4 shows the results achieved by the TECLA for a validation with 6 folders executed 10 times. The values in bold are the best average accuracy and F-measure results obtained by the classifiers for each temperature.

For the artisan temperament, the KNN algorithm with K = 1 obtained an average accuracy of 80.44% and an F-measure of 88.91%. With a better performance, that is, a greater number of correctly labeled objects, it was the SVM algorithm with an average accuracy of 88.37%, followed by the Random Forest with an average accuracy of 87.95%. The SVM presented a higher average accuracy and a 100% recall, whilst the Random Forest presented a better precision than SVM.

In relation to the guardian temperament, the most assertive prediction was by the SVM algorithm, with an average accuracy of 86.92% and F-measure of 93%, followed by the Random Forest with an average accuracy of 86.32%. The lowest average accuracy (78.36%) was for the KNN with K = 1.

The SVM also performed better for the idealist and rational temperaments. The idealist temperament had an average accuracy of 55.61% and F-measure of 71.46% and the rational temperament had an average accuracy of 69.09% and F-measure of 81.71%. It is concluded that these two temperaments had a good performance in the labeling of objects due to the average accuracy being greater than 50%.

In general, the SVM obtained better accuracy for all temperaments, but for the artisan and guardian temperaments the Random Forests presented very close average accuracies to the SVM.

2) TF-IDF

Table 5 shows the results achieved by the TECLA framework for a cross-validation with 10 folders, executed 10 times. The values in bold are the best average accuracy and F-measure results obtained by the binary classifiers for each temperature.

The artisan temperament obtained a high average accuracy (90.93%) with the KNN, K = 3 and F-measure of 95,09% which makes effective the result obtained by the KNN effective, K = 3, followed by the SVM with an average accuracy of 88.35%. The SVM had the highest average accuracy for the guardian temperament, but in this case the KNN for K = 3 had a very low performance. One hypothesis for this low value is the imbalance of the database, so when increasing the number of neighbors the algorithm can not label the object. For the idealist temperament the SVM and KNN (K = 3) were practi cally even. For the rational temper-ament the best performance was for the KNN algorithm with K = 2, with an average accuracy of

Table 4: Accuracy (Acc), Precision (Pre), Recall (Rec) and F-measure (M-F) for the four temperaments using 6 folders and 10 iterations.

	LIWC	1NN	2NN	3NN	Random Forest	SVM
Artisan	Acc	**80,44% ± 0,71%**	87,62% ± 0,37%	87,62% ± 0,37%	**87,95% ± 0,16%**	**88,37% ± 0,00%**
	Pre	88,79% ± 0,14%	88,47% ± 0,07%	88,47% ± 0,07%	88,41% ± 0,06%	88,37% ± 0,00%
	Rec	89,10% ± 0,87%	98,86% ± 0,44%	98,86% ± 0,44%	99,39% ± 0,15%	100,00% ± 0,00%
	M-F	88,91% ± 0,47%	93,37% ± 0,23%	93,37% ± 0,23%	**93,58% ± 0,09%**	**93,82% ± 0,00%**
Guardian	Acc	**78,36 ± 0,62%**	85,67% ± 0,10%	85,74% ± 0,10%	**86,32% ± 0,11%**	**86,92% ± 0,01%**
	Pre	87,05% ± 0,07%	86,94% ± 0,04%	86,94% ± 0,04%	87,03% ± 0,06%	86,92% ± 0,00%
	Rec	88,22% ± 0,76%	98,27% ± 0,09%	98,36% ± 0,09%	99,02% ± 0,11%	100,00% ± 0,01%
	M-F	87,61% ± 0,43%	92,25% ± 0,06%	92,30% ± 0,06%	92,63% ± 0,06%	**93,00% ± 0,01%**
Idealist	Acc	54,97% ± 0,46%	54,97% ± 0,46%	**52,57% ± 0,61%**	54,27% ± 0,40%	**55,61% ± 0,01%**
	Pre	56,80% ± 0,27%	56,80% ± 0,27%	57,88% ± 0,57%	56,67% ± 0,26%	55,61% ± 0,01%
	Rec	79,44% ± 0,80%	79,44% ± 0,80%	54,18% ± 0,97%	75,65% ± 1,04%	100,00% ± 0,00%
	M-F	66,19% ± 0,33%	66,19% ± 0,33%	55,86% ± 0,67%	64,76% ± 0,49%	**71,46% ± 0,02%**
Rational	Acc	**59,12% ± 0,58%**	65,82% ± 0,49%	87,62% ± 0,37%	66,62% ± 0,26%	**69,09% ± 0,03%**
	Pre	69,72% ± 0,21%	69,27% ± 0,13%	88,47% ± 0,07%	69,74% ± 0,14%	69,10% ± 0,01%
	Rec	72,17% ± 1,20%	90,84% ± 1,16%	98,86% ± 0,44%	91,38% ± 0,40%	99,97% ± 0,04%
	M-F	70,85% ± 0,65%	78,57% ± 0,47%	74,78% ± 0,27%	79,09% ± 0,19%	**81,71% ± 0,03%**

Table 5: Accuracy (Acc), Precision (Pre), Recall (Rec) and F-measure (M-F) for the four temperaments using 10 folders and 10 iterations.

	TF-IDF	**1NN**	**2NN**	**3NN**	**Random Forest**	**SVM**
Artisan	Acc	87,31% ±3,01%	90,25% ± 0,09%	**90,93% ± 0,06%**	87,69% ± 0,07%	**88,35% ± 0,07%**
	Pre	87,31% ± 3,01%	90,25% ± 0,09%	90,93% ± 0,06%	87,69% ± 0,07%	88,35% ± 0,07%
	Rec	99,00% ± 3,00%	100,00% ± 0,00%	100,00% ±0,00%	100,00% ± 0,00%	100,00% ± 0,00%
	M-F	92,56% ± 2,99%	94,73% ± 0,07%	**95,09% ±0,06%**	93,25% ± 0,08%	**93,60% ± 0,07%**
Guardian	Acc	88,98% ± 0,06%	85,08% ± 0,09%	**27,13% ± 3,73%**	88,25% ± 0,07%	**90,93% ± 0,08%**
	Pre	88,98% ± 0,06%	85,08% ± 0,09%	16,27% ± 3,25%	88,25% ± 0,07%	90,93% ± 0,08%
	Rec	100,00% ± 0,00%	100,00% ± 0,00%	19,00% ± 3,00%	100,00% ± 0,00%	100,00% ± 0,00%
	M-F	94,04% ± 0,05%	91,68% ± 0,11%	17,50% ± 3,16%	93,55% ± 0,13%	**95,15% ± 0,06%**
Idealist	Acc	52,40% ± 2,14%	51,98% ± 0,10%	**61,05% ± 0,08%**	54,42% ± 3,29%	**61,04% ± 0,12%**
	Pre	48,14% ± 1,05%	51,98% ± 0,10%	61,05% ± 0,08%	50,12% ± 5,65%	61,04% ± 0,12%
	Rec	90,00% ± 0,00%	100,00% ± 0,00%	100,00% ± 0,00%	88,00% ± 8,72%	100,00% ± 0,00%
	M-F	62,12% ± 0,99%	67,74% ± 0,20%	**75,13% ± 0,27%**	**63,32% ± 6,78%**	**75,08% ± 0,34%**
Rational	Acc	65,52% ± 1,43%	**71,42% ± 0,10%**	70,80% ± 0,11%	65,60% ± 1,82%	68,21% ± 0,06%
	Pre	62,18% ± 0,72%	71,42% ±0,10%	70,80% ± 0,11%	65,40% ± 2,41%	68,21% ± 0,06%
	Rec	90,00% ± 0,00%	100,00% ±0,00%	100,00% ± 0,00%	99,00% ± 3,00%	100,00% ± 0,00%
	M-F	73,16% ± 0,50%	**82,85% ±0,27%**	82,51% ± 0,27%	78,21% ± 2,63%	80,70% ± 0,013%

82.85%. Again the SVM was the algorithm that presented the best average performance among the evaluated ones.

5 CONCLUSIONS AND FUTURE WORK

Temperament influences the way we perceive and react to the world. Understanding temperament is of crucial importance to our lives and to position ourselves properly in the market. Normally, one's temperament can be known by filling in tests, such as the MBTI (Myers-Briggs Type Indicator). The hypothesis of this research is that it is possible to identify the temperament of a person in a passive way, only by using data obtained from the social media of the person. For this, a database of tweets containing the MBTI result of Twitter users was employed. These data were used to generate predictive models of temperament.

The documents (Tweets) were structured with the Portuguese dictionary LIWC that groups words into categories. The calculation of the frequency of words was carried out to show which category is most talked about by artisan, guardian, idealistic and rational temperaments. In this analysis it is possible to identify the writing tendency of the user associated with the subject that is most identified, perception among others.The tweets were structured using LIWC and TF-IDF. For classification via LIWC the best accuracy results were achieved for the artisan and guardian temperaments trained with SVM. For the TF-IDF the highest average accuracy was for the artisan, guardian and idealist temperaments, also with emphasis on the SVM algorithm. For the representation using the TF-IDF the best average accuracy was observed for the artisan and guardian temperaments for the KNN (K = 3) and SVM algorithms.

As a future work, we intend to carry out a case study using the TECLA framework with a database composed of a set of volunteer users who will answer the MBTI test form and share their social profiles so that we can use their data to train the TECLA framework and classify temperament.

Another improvement to be made is the study of the content of the documents to investigate why the classifiers have low accuracy and how much the unbalanced basis interferes in this result to verifying whether there is need for treatment for the unbalanced classes.

ACKNOWLEDGEMENTS

The authors thank Mackenzie University, Mackpesquisa, CNPq, Capes and FAPESP for the financial support.

REFERENCES

Calegari, M. d., & Gemignani, O. H. (2006). *Temperamento e Carreira.* São Paulo: Summus Editorial.

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced approaches in Analyzing Unstructured Data.* Cambridge university press.

Gundecha, P., & Liu, H. (2012). Mining social media: a brief introduction. *New Directions in Informatics, Optimization, Logistics, and Production. Informs*, pp. 1-17.

Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science, 17*, 26-32.

Hall, C. S., Lindzey, G., & Campbell, J. B. (2000). *Teorias da Personalidade.* Porto Alegre: Artmed.

Ito, P. d., & Guzzo, R. S. (2002). Diferenças individuais: temperamento e personalidade; importância da teoria. *Estudos de Psicologia*, pp. 91-100.

Keirsey, D. (1998). *Please Undestand Me II: Temperament, Character, Intelligence.* Prometheus Nemesis Book Company.

Keirsey, D. M. (1996). *Keirsey.com*. (Corporate Offices) Acesso em 12/10/2017 de 10 de 2017, disponível em ://www.keirsey.com/4temps/overview_temperaments.asp

Komisin, M. C., & Guinn, C. I. (2012). Identifying personality types using document classification methods. *In: FLAIRS Conference.*

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th international conference on World wide web. ACM.*, 591-600.

Lima, A. C. (2016). *Mineração de Mídias Sociais como Ferramenta para a Análise da Tríade da Persona Virtual.* São Paulo.

Lima, A. C., & de Castro, L. N. (2016). Predicting Temperament from Twitter Data. *Advanced Applied Informatics (IIAI-AAI), 2016 5th IIAI International Congress on. IEEE*.

Nor Rahayu, N., Zainol, Z., & Yoong, T. L. (2016). A comparative study of different classifiers for automatic personality prediction. *Control System, Computing and Engineering (ICCSCE), 2016 6th IEEE International Conference on. IEE*, 435-440.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Perro, M. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12.Oct, 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.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015.*

Plank, B., & Dirk, H. (2015). Personality Traits on Twitter-or-How to Get 1, 500 Personality Tests in a Week. *Proceedigs of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis(WASSA 2015)*, pp. 92-98.

Spencer, J., & Uchyigit, G. (2012). Sentimentor: Sentiment analysis of twitter data. *Proceedings of European conference on machine learning and principles and practice of knowledge discovery in databases*.

Verhoeven, B., Daelemans, W., & Plank, B. (2016). Twisty: A Multilingual Twitter Stylometry Corpus for Gender and Personality Profiling. *Proceedings of the 10th International Conference on Language Resources and Evaluation*.

Xavier, O. C., & Carvalho, C. L. (2011). *Desenvolvimento de Aplicações Sociais A Partir de APIs em Redes Sociais Online*. UFG. Goiânia.