Natural Language Processing Approach for Classification of Archetypes Using Text on Business Environments

Richard Vinícius Rezende Mariano[1], Ana Carolina Conceição de Jesus[1], Alessandro Garcia Vieira[2], Jessica da Assunção Almeida de Lima[2], Giulia Zanon de Castro[2] and Wladmir Cardoso Brandão[1,2]

[1]*IRIS Research Laboratory, Department of Computer Science, Pontifical Catholic University of Minas Gerais (PUC Minas), Belo Horizonte, MG, Brazil*

[2]*Data Science Laboratory (SOLAB), Sólides S.A., Belo Horizonte, MG, Brazil*

Keywords: People Analytics, Text Classification, Behavioral Classification, Natural Language Processing, Machine Learning, Support Vector Machine.

Abstract: Organizations increasingly offer resources to improve performance, minimize costs, and achieve better results. An organization is the individuals who work or provide services in it. Therefore, good organizational performance directly results from the good work of its collaborators. Identifying the archetype in the business environment can combine individuals with companies, which can improve the organizational environment and enhance the development of the individual. A person leaves traces of his behavior in what he produces, such as videos and texts. Some studies point to the possibility of identifying a behavioral profile from a textual production. In this work, we seek to identify the archetype of individuals within the business environment based on their curriculum texts. We combine the behavioral profile assessment (BPA) archetypes (Planner, Analyst, Communicator, and Executor) with 26,636 curriculum to apply machine learning models. For this task, we used classification and regression approaches. The main algorithm used for the approaches was the SVM. The results suggest that the archetypes are better modeled using regression techniques, obtaining an MSE of 4.49 in the best case. We also provide a visual explanation example to understand the model outputs.

1 INTRODUCTION

The study of behavioral profiles, also called archetypes, is a common practice in psychology. This study defines a group based on behavior patterns, communication style, and reactions to the environment and people. Understanding a person's archetype can help them better understand themselves and their actions, in personal, family, and professional environments. In the professional context, companies are increasingly using psychological theories and technology to make decisions about their workforce. Identifying a company's needs and the best profile for them is one of the main focuses of HR teams.

Having the right professionals in the right companies allows for greater efficiency in the job market. Companies can benefit by placing employees with specific behavioral profiles in demanding tasks, hiring based on needs, assembling teams focused on a particular job, or possessing a combination of skills to achieve the result. Additionally, understanding employees' behavior profiles helps companies effectively deal with any difficulties they may encounter, overcome problems, reduce unnecessary turnover, and facilitate the growth of individuals and the company.

The employee also has gains, avoiding entering companies that do not understand their needs and facilitating entry into companies that fit. Being in the right company creates more meaningful opportunities for personal and professional growth.

Several behavioral classification tools have emerged from studying psychological and behavioral profiles. They are focused mainly on Eysenck Factors (Eysenck and Eysenck, 1965), *DISC* (Marston, 1928) and *BigFive* (McDougall, 1932) models, the last being the most common. These models are widely used in the literature to explore the classification of psychological profiles.

Within this scope, we raise the question: "Does an individual transmit their behavior, profile, and archetype in their texts?" Psychology points out a correlation between personality traits and linguistic level, including acoustic parameters (Smith et al., 1975) and lexical category (Pennebaker et al., 2003). We believe that each person leaves their mark, writing style, and personality in their textual production.

In this work, we will expand the studies of iden-

Mariano, R., Conceição de Jesus, A., Vieira, A., Almeida de Lima, J., Zanon de Castro, G. and Brandão, W.
Natural Language Processing Approach for Classification of Archetypes Using Text on Business Environments.
DOI: 10.5220/0011856200003467
In *Proceedings of the 25th International Conference on Enterprise Information Systems (ICEIS 2023) - Volume 1*, pages 501-508
ISBN: 978-989-758-648-4; ISSN: 2184-4992

tification and classification of the behavioral profile, focusing on the organizational environment. More specifically, in the Brazilian business environment. With these studies, we raise the following hypothesis.

> *Hypothesis:* Psychological and behavioral profiles within the organizational environment can be identified from textual productions.

To evaluate this hypothesis, we will combine two main Natural Language Processing (NLP) techniques, vector representation of texts and characteristics extraction. Together, they bring much information about the text, which can be studied, understood, and used to construct a classification model. We aim to build this model and apply it in a behavioral assessment aimed at the corporate environment. For this work, we chose to use a Behavioral Profiler Assessment (BPA), built with a direct focus on the organizational environment. Since this tool focuses on Brazilian business culture, we chose to use texts in Portuguese extracted from the curriculum. These resumes are diverse and have been collected from multiple companies and people from different places. The main contributions of this work are:

- We propose a methodology for building a behavioral profile prediction model using textual data from candidates' CVs;
- We provide an explainability analysis of the model outputs, helping to understand the textual patterns of different behavioral profiles.

This work is divided as follows. Section 2 defines the background to understanding this research. In Section 3, we present the related works. The methodology is described in the Section 4 and Section 5 displays the experiments and results. Finally, Section 6 concludes our work and presents the future works.

2 THEORETICAL BACKGROUND

2.1 People Analytics

People Analytics refers to collecting, organizing, and utilizing people's data, usually in a business environment, to help people management. This methodology has become increasingly present with HR teams adopting new technologies (Raguvir and Babu, 2020). The main focus is identifying behavior information that may be used to track conduct, performance, and results. The applications of People Analytics are diverse. In the business context, its primary focus is to increase efficiency and productivity, reduce conflicts, and create a better work environment.

2.2 Behavioral Study

Human behavior is something of great fascination for humanity. Across different times, places, and cultures, people have attempted to categorize individuals based on their characteristics and behavior into distinct groups. These groups are commonly referred to as the behavioral profile or archetype. By understanding the profilers, we can better understand how a person fits into society, their potential strengths and weaknesses, and the impact they can have.

The number of behavioral profiles grouped throughout history varied mainly between four/ five personalities. For example, the prophet Ezekiel saw humans as four personalities (lion, ox, man, eagle), while the Greeks linked human behavior to the four elements of nature (fire, water, air, earth). Hippocrates, the father of Western medicine, proposes that the human temperament is directly related to the balance of the essential bodily fluids (blood, black bile, yellow bile, and phlegm), refer to happy, somber, enthusiastic, and calm temperaments, respectively.

The relationship between human behavior and nature with elements of nature and body parts is also found in Chinese culture. Each element is associated with a specific personality type and represented by an organ in the body: water represents the kidney, wood the liver, fire the heart, earth the pancreas, and metal represents the lung. This concept is based on Traditional Chinese Medicine.

The psychiatrist Carl Gustav Jung brings one of the most well-known classifications of individuals into four groups: feeling, sensation, intuition, and thinking (Jung and Hull, 1971). In the early 20th century, American psychologist William Moulton Marston created the DISC methodology, which outlines four main behavioral types: dominance (control, power, and assertiveness), influence (communication and social relationships), stability (patience and persistence), and caution (organization and structure) (Marston, 1928). Additionally, McDougall proposed the BigFive model in the 20th century, which defines five main factors influencing personality: neuroticism, extroversion, pleasantness, conscientiousness, and openness to experience (McDougall, 1932).

Despite the multiple approaches to studying human behavior, dating from different times and cultures, we can observe a relationship between them and a constant common desire to understand each other.

3 RELATED WORK

Several studies focus on the recognition of personality based on the BigFive model. One of the pioneers is (Pennebaker and King, 1999), with a focus on analyzing the reliability of its feature extraction techniques. The authors propose a tool called LIWC. (Oberlander and Nowson, 2006; Nowson and Oberlander, 2007) used n-gram techniques, specifically bi-grams and tri-grams, along with binary and multi–class classification to measure accuracy based on the BigFive model. They chose to work with only 4 out of 5 BigFive profiles, leaving the Openness profile aside.

Following this line, using the BigFive, several approaches are made through a binary classification technique (Argamon et al., 2005; Mairesse et al., 2007; Sumner et al., 2012; Park et al., 2014; Majumder et al., 2017; Santos et al., 2017; Vu et al., 2017; dos Santos and Paraboni, 2019). In most cases, accuracy is used as a metric for evaluation. (Argamon et al., 2005) uses *f1-score* in conjunction with accuracy and uses a Grammar Parser as an attribute of its data. Argamon focused his experiments on Neuroticism and Extraversion personalities, applying binary classifiers. (Mairesse et al., 2007) extracts attributes from feature extraction, and in addition to binary classification, it also uses ranking and regression techniques. Their ranking results ranged between 56% and 63% on written text data and 61% on spoken data.

(Majumder et al., 2017) uses word embeddings in addition to feature extraction, using CNN networks for its binary classification. Its biggest result was in the openness profile with 62.7% accuracy. Approaches using Word-net and SentiWordNet are explored by (Vu et al., 2017), they used data extracted from social networks and obtained best results in 3 of the 5 profiles in relation to Majumder's work. (Santos et al., 2017) and (dos Santos and Paraboni, 2019) also evaluate feature extraction using the f1-score as a metric. The authors evaluate the BigFive model by applying NLP techniques such as Bag of Words and SkipGram, building 6 different textual datasets.

Others techniques is also explore in literature, such as the use of regression (Gill et al., 2009), (Golbeck et al., 2011), (Karanatsiou et al., 2022), using three or more classes in their models. (Karanatsiou et al., 2022) combines Bag–of–words with Post-tagging and emotion extraction in its models, using RMSE and MAE as metrics, to calculate the error of the regressor models. Other personality models are also used in the literature for automatic recognition. Like the Eysenck Factors (Gill and Oberlander, 2002), the MBTI typology (Luyckx and Daelemans, 2008), and the DISC model (Pereira, 2021).

Although some papers follow other ways, we can notice a great concentration on using of the BigFive model as a personality cataloging techniques and a preference for dividing the problem into minors binary classification Despite this, we can see an evolution in the research area (Eisenack et al., 2021).

4 METHODOLOGY

4.1 Behavioral Profile Assessment

The behavioral mapping tool used as the basis for creating the dataset for this work is a built based on 8 methodologies for mapping behavioral profiles, with methodologies from different times and places. The methodology divides the profiles into 4 (Analyst, Communicator, Executor, and Planner) and delivers a percentage referring to each profile, where the sum of the percentages is equal to 100. Thus, an individual with a certain archetype is considered, if the percentage referring to that archetype is equal to or greater than 25%. An example can be seen in Figure 1, the individual is considered a Communicator Executor, since he has both archetypes above 25%, being Executor his main archetype. The BPA approach allows for the possibility of various combinations and levels, which makes each personality unique. A brief explanation of each BPA archetype is described below.

- Analyst: Detailed, rigid and calm. With discreet and observant behavior, they are very detail-oriented, but have a lot of focus, intelligence and perfectionism. They have ease with the field of the arts, but they charge a lot, they are skilled with detailed tasks or risk management.
- Communicator: They are outgoing, talkative and active. They adapt easily, have ease in communication, like jobs that involve movement and autonomy. They work best as a team, are festive, lively and relaxed, are imaginative and artistic.
- Executor: Active, dynamic and competitive. Not afraid to take risks and face challenges. They have leadership characteristics, are self-confident, have autonomy and independence. Their Reasoning tends to be more logical and deductive, they appreciate challenges and obstacles, tend to execute before thinking.
- Planner: Calm and prudent. They like routine, and to act with common sense, following norms and rules. Generally introverted, but easy to get along with. They are patient and observant, act with tranquility and discipline.

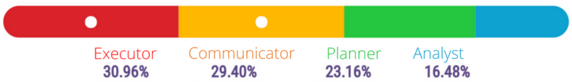

Figure 1: BPA report.

4.2 The Dataset

The dataset used in this work is a private base, extracted from the BPA tool. This base consists of 26636 instances. Each instance consists of a text written by an individual, the respective percentages referring to each archetype of that individual, and the formation of their final profile. The classes composition of the dataset is divided as follows, 38.5% have the Analyst profile above or equal to 25%. 50.81% have the Communicator profile, 58.67% Executor, and finally, 51.96% have the Planner profile. Remembering that each individual can have 1 to 3 profiles, it is enough that their percentage in that profile is above or equal to 25%. The complete composition of the dataset following the number of instances for each possible combination can be seen in Table 1, where A means Analyst, C refers to Communicator, E to Executor, and finally, P means Planner.

Table 1: Distribution of Archetypes.

Main Analyst		Main Communicator	
A	478	C	974
AC	245	CA	297
AE	686	CE	3274
AP	2461	CP	1247
ACE	30	CAE	27
ACP	75	CAP	53
AEC	37	CEA	46
AEP	243	CEP	326
APC	120	CPA	93
APE	300	CPE	238
Main Executor		**Main Planner**	
E	1671	P	882
EA	1004	PA	2731
EC	3963	PC	1165
EP	1122	PE	876
EAC	63	PAC	164
EAP	171	PAE	209
ECA	91	PCA	139
ECP	299	PCE	162
EPA	185	PEA	188
EPC	182	PEC	114

4.3 Features

4.3.1 Text-Vector

There are several ways to represent the text through vectors of words, which will then be used to train a learning model. From basic TF-IDF to more complex techniques like word embeddings.

After tested some techniques, we chose the one that performed best, tokenization. In this representation, each word in every base has its representation in number, so each text has its vector representation of numbers in a unique way, then we apply a pad sequence that leaves all vectors with the same size.

Preprocessing: To work with vectors of words, it is first necessary to clear this data to facilitate representation, and also facilitate classification learning. It is necessary to be very careful with the pre-processing because pre-processing will not always help to solve a problem. So it is necessary to do several experiments, adding and removing to see how the model performs.

The pre-processing done in this work are: Remove special characters, punctuation and accentuation; Remove stopwords, the words the most common in a language; Lower all text; Lemmatization and stemming. Grouping the inflected forms of a word so that they can be analyzed as a single item.

4.3.2 Characteristics Extraction

Extracting the characteristics of a text allows greater exploration of what is being said by the the writer. The idea is to go beyond the text and obtain information about its composition. For this, we use a post-tag tool as an aid. With this tool, we will extract the number of times the text has each grammatical class.

To level the data, we will also extract the number of words per texts. Doing the proportion of each grammatical class in relation to the total number of words, then obtaining the percentage of representation of that grammatical class in the texts.

The Post-tagger tool used in this work is open-source and available on Github[1]. The tool was pre-trained to handle sentences in Portuguese and reaches up to 92.2% accuracy when tagging texts. In the end, the features consists of the number of words per text plus the following parts of speech: adjective, adverb, article, conjunction, interjection, noun, proper noun, number, participle, pronoun, preposition, and verb. With a total of 14 features.

4.4 Approaches

4.4.1 Multi–Class Classification Approach

By viewing the problem as a multi–class problem each instance can have from 1 to 4 classes. The most common being having 2 of the 4 classes, which occurs 72.21% of the times in the dataset, while 14.45% have only one class and 13.34% have 3 classes.

[1]https://github.com/inoueMashuu/POS-tagger-portuguese-nltk

Although the instance is multi–classes as a result of an individual being able to possess more than one archetype, the highest percentage archetype can be considered its "main archetype". Following this reasoning, the multi–class approach consists of training the learning algorithm based on the main archetype and using the output probabilities to verify the performance of the classifier. To analyze this classification, we divided the problem into 4 scenarios of analyse, so that it is possible to observe different aspects of the behavior and performance of the classification.

Analyse 1 (A1): Hit only the Main Archetype. If the highest probability in the classifier output is equivalent to the Main Archetype of the instance.

Analyse 2 (A2): Main Archetype Probability above 25%. If the output probability of the Main Archetype classifier is equal to or above 25% it is a hit, even if there is another archetype with a higher output probability.

Analyse 3 (A3): If any probability of the classifier equal to or above 25% is equivalent to some archetype of the individual. In this scenario, the highest probability of the classifier, or the highest percentage of the instance archetype, does not matter.

Analyse 4 (A4): Each archetype is considered a hit or miss. This analyse brings more reliability to the result. The classifier probability of each archetype is compared with the percentage of each archetype of the instance. That is, for each instance we have a total of 4 hits or misses. The hit is considered when the classifier percentage is equal to or above 25% and the instance has that archetype, but also when the classifier percentage is below 25% and the instance does not have that archetype.

It is worth remembering that the BPA tool that defines the individual's profile, uses the threshold of 25% to define the individual's archetypes, and therefore, we decided to use this threshold in our scenarios for the experiment.

4.4.2 Binary Classification Approach

The binary classification allows us to divide the problem into four smaller problems. Assigning each task to a different classifier, and each classifier working on the prediction of a single archetype. The idea with this approach is to achieve 3 main goals. i) The ability to compare performance with other works, since many papers in the literature used binary classification by profile. ii)Analyze the performance of machine learning methods in the simplified classification, which allows better adjustment of parameters and metrics to solve the problem. iii) It allows a better analysis of the decision making of the algorithms, which will allow us a greater explainability of the models.

We chose to use accuracy and f1–score for evaluation metrics. Below is an explanation of each metric.

a. *Accuracy:* Expresses the number of model hits in relation to the total number of samples.

b. *F1–score:* The average of accuracy with the number of hits per number of predictions by class.

The algorithms that we will use in this classification, in both evaluation of the grammatical features extracted with post-tagger, and the word vector features, is the SVM. The SVM algorithm is widely used in the literature, and fits our problem. It is simple and efficient, especially in classifying binary problems.

4.4.3 Regression Approach

Regression algorithms allow us to use continuous data for training and prediction. This approach allows us to work directly with the percentages passed by the BPA. In that case, we will also build four different regressors, one for each archetype. These regressors will be made using the SVM algorithm.

In this approach, what matters is the difference between the right answer and what was predicted, that is, the error. We then chose two techniques for error calculation, and both metrics calculate the distance between actual values and predictions. The first is RMSE (root mean squared error) squares the distance for each instance before calculating the average, this metric suffers from data where there are many outliers. The second is MAE (mean absolut error), that calculates exactly the average of the distances between actual values and predictions. For both error metrics, the smaller the value, better is the results.

4.5 Interpretability

The goal of interpretability is to understand the reasons that made a machine learning algorithm makes a decision. Machine learning algorithms tend to be, in general, a "black box". Where in the end, we only extract some metrics such as accuracy and f1–score, without understanding the reasons behind the predictions. In simpler classifiers, we can come to understand the path taken by the algorithm, such as in the case of decision trees. But in more complex cases, such as neural networks, the path is foggy, due to a large number of parameters, which can be thousands or even millions, understanding cannot be done quickly, which prevents quick decision-making.

To help solve this problem, interpretability techniques can be used. In this context, we have LIME, a method of local surrogate models. The objective of this model is to approximate the results of the black box models, however, focused on local training, thus being able to explain individual predictions.

5 EXPERIMENTS AND RESULTS

5.1 Multi–Class Classification Report

This multi–class classification approach allows us an initial overview in the analysis of the problem. We use the SVM algorithm, the evaluation metrics are defined in section 4.4.1, and the results can be seen in Table 2. We can notice that the A1 and A2 analyzes are limited, since we are considering only one archetype in the evaluation, and the individual has a little of each archetype. The A3 assessment is positive, but it is not very reliable, its metrics tend to be correct even if randomly. The A4 is a good metric to evaluate, as it considers the hit and error in the four archetypes, getting closer to the reality delivered by the BPA.

Table 2: Multi–class Classification report.

	A1	A2	A3	A4
Hit	0.33	0.54	0.79	0.55

The challenge of this approach is that although one archetype stands out over the others, the individual has a little of each archetype, even having more than one dominant profile. Thinking about it, we took the path of making a binary classification, which allows an analysis of each profile separately.

5.2 Binary Classification Results

We then have four classifiers, each focused on classifying one of the archetypes in the database. In this way, we generated four datasets derived from the main dataset, considering that some instances have more than one archetype, some data can be repeated, but this does not affect the models, since the classifiers are independent. We apply data balancing to each of these four datasets as needed. For example, we have more executors than non-executors, so we decrease the number of executors in the base.

We apply the svm algorithm to classify both sets of features. The metrics used was accuracy and f1–score, the results can be seen in Table 3.

Experiments with text vectors were better than Post-tagger, in all aspects. The SVM algorithm combine with Text vector representation brought accuracy above or equal to 63% for all 4 archetypes, standing out mainly with Planners, with an accuracy of 65%. The Post-tagger approach showed little relevant results in terms of accuracy and f1–score. We believe

Table 3: Binary Classification report.

	Text Vector		Pos-tagger	
	Accuracy	f–score	Accuracy	f–score
P	0.65	0.62	0.53	0.52
A	0.63	0.63	0.51	0.51
C	0.63	0.60	0.51	0.51
E	0.63	0.63	0.52	0.52

Figure 2: Real sample LIME report.

that a possible combination of both techniques, text vector and post-tagger, can bring considerable improvements in the construction of a model.

5.2.1 Explainability

In this section, we provide the local explainability some sample texts. First, we searched the BPA for the main words that describe each profile, these words can be seen in table 4.

Table 4: Archetypes main description words.

Planner	Analyst	Communicator	Executor
Calm	Calm	Active	Active
Observer	Observer	Extrovert	Competitive
Disciplined	Disciplined	Speakers	Leader
Quiet	Discreet	Communicative	Determined
Introverts	Organized	Independence	Independence
Routine	Transparent	Sociable	Persistent
Reliable	Honest	Empathic	Logical
Patients	Detail	Persuasive	Self-confident
Righteous	Perfectionists	Optimistic	Intuitive
Flexible	Thoughtful	charismatic	Disposed

Let's then explore some samples of local explainability, more specifically, two examples. One sample extracted directly from the dataset, and another text created by us seeking to explore the model's decisions. In this analysis we will use the binary models of Executor classification. First, we analyzing the sample taken from the dataset, we apply LIME explicability as we can see in Figure 2, the full text will not be displayed for privacy reasons.

The explanation of the figure shows the local ex-

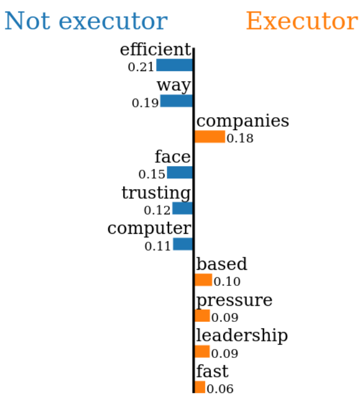

Figure 3: Example sample LIME report.

plainability of a curriculum text. On the right, we have the features that have a positive correlation with the output of the analyzed class, and on the left a negative correlation. For example, the word "executive" is the word that has the highest correlation with the class to be predicted. While the words "young" and "agility" are the main words with opposite correlation to the analyzed class.

Now we will apply the local explainability to a text created by us, just for research purposes, the result can be see in Figure 3. And the text follows.

> I'm 23 years old, graduated in Computer Science and have experience in software development. I'm looking for a job where I can demonstrate my qualities, take risks and face challenges. I am an independent person, able to solve problems under pressure and in a practical way. Relationships with co–workers in previous companies were mainly based on competitiveness. Among my main qualities, I am trusting, proactive, persistent and have leadership skills, I like to do my tasks fast and efficient. I would say that my main defect is to be inflexible in my ideals.

We can notice that in Table 4, words like "active" and "leader" are used to describe the Executor archetype, as well as in Figure 3 that words like "fast" and "leadership" has a positive correlation with the class. However, some words like "efficient" were negatively correlated with the Executor profile. These variations occur because each person is formed by the 4 archetypes. Therefore, we decided to explore other approach to archetype inference, using regression.

5.3 Regression Report

Regression experiments were done with the SVM algorithm for each of the four archetypes, then extracting the error, the error metrics used were RMSE and MAE. The lower the value of these metrics, the more efficient the model is. The results obtained can be seen in table 5, we note that the MAE metric has better results than the RMSE, which points to the existence of some outliers that generate an increase in the RMSE. The results obtained are positive and show that the prediction is close to the true label. The Regression obtained better results when compared to the classification approach.

Table 5: Error Regression report.

	RMSE	MAE
Planner	5.98	4.49
Analyst	7.02	5.24
Communicator	7.06	5.36
Executor	7.05	5.24

6 CONCLUSIONS

Exploring behavioral profiles is a relevant task to generate improvements in the People Analytics area. Placing the right people in the right companies brings more efficiency and harmony.

In the present paper, we proposed approaches to identify the archetype from textual productions automatically. In particular, we proposed using a database with a new methodology focused on the business area, precisely the Brazilian business scenario.

Our experiments showed potential in the classification of archetypes. We use two representations type of features, direct use of the text, transforming it into a vector, and information extraction from the text, in this case, the distribution of grammatical classes.

First, make a multi–class approach, trying to predict all profiles with a single model, depending on how we evaluate this approach, the results are interesting, but below expectations. We then proceeded to a binary approach, building a model for each archetype. This approach proved to be better, mainly in the combination of SVM with text vector, obtaining accuracy and f1–score at least 63% for all profiles, with emphasis on planner with an accuracy of 65%. Still, in this binary approach, we apply interpretability techniques, to explore the decisions made by the models, and bring more transparency to the problem.

The last approach was using regression to identify the archetypes and calculating the error through two metrics, RMSE and MAE, emphasizing the planner with an RMSE of 5.98 and an MAE of only 4.49.

When it comes to applications, we believe that this type of behavioral analysis from texts can add to the selection of people by companies. But it is still early to say that it can replace other selection processes, the human factor is still very important and cannot be removed. The idea is to give one more tool option to be used, which allows more possibilities to find the best match between company and the candidate.

In future work, we intend to expand our features, increase the number of characteristics extracted, and explore new vector text representation to improve our results. Furthermore, regression techniques show more promise than classification techniques, so we want to explore this type of model further.

ACKNOWLEDGEMENTS

The present work was carried out with the support of Sólides S.A. The authors thank the partial support of the Pontifical Catholic University of Minas Gerais (PUC Minas).

REFERENCES

Argamon, S., S, D., Koppel, M., and Pennebaker, J. (2005). Lexical predictors of personality type.

dos Santos, W. R. and Paraboni, I. (2019). Personality facets recognition from text. *ArXiv*, abs/1810.02980.

Eisenack, K., Oberlack, C., and Sietz, D. (2021). Avenues of archetype analysis: Roots, achievements and next steps in sustainability research. *ECOLOGY AND SOCIETY*, 26.

Eysenck, H. J. and Eysenck, S. (1965). The eysenck personality inventory. *British Journal of Educational Studies*, 14(1).

Gill, A., Nowson, S., and Oberlander, J. (2009). What are they blogging about? personality, topic and motivation in blogs.

Gill, A. and Oberlander, J. (2002). Taking care of the linguistic features of extraversion. In Gray, W. and Schunn, C., editors, *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 363–368. Lawrence Erlbaum Associates. 24th Annual Conference of the Cognitive Science Society ; Conference date: 07-08-2002 Through 10-08-2002.

Golbeck, J., Robles, C., Edmondson, M., and Turner, K. (2011). Predicting personality from twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 149–156.

Jung, C. G. and Hull, R. F. C. (1971). *Psychological types*. Number 6 in Bollingen series. Routledge, London.

Karanatsiou, D., Sermpezis, P., Gruda, D., Kafetsios, K., Dimitriadis, I., and Vakali, A. (2022). My tweets bring all the traits to the yard: Predicting personality and relational traits in online social networks. *ACM Trans. Web*, 16(2).

Luyckx, K. and Daelemans, W. (2008). *Personae: a Corpus for Author and Personality Prediction from Text*.

Mairesse, F., Walker, M. A., Mehl, M. R., and Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Int. Res.*, 30(1):457–500.

Majumder, N., Poria, S., Gelbukh, A., and Cambria, E. (2017). Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79.

Marston, W. (1928). *Emotions of Normal People*. International library of psychology, philosophy, and scientific method. K. Paul, Trench, Trubner & Company Limited.

McDougall, W. (1932). Of The Words Character and Personality. *Journal of Personality*, Vol. 1(1):3–16.

Nowson, S. and Oberlander, J. (2007). Identifying more bloggers: Towards large scale personality classification of personal weblogs. In *ICWSM*.

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

Park, G., Schwartz, H., Eichstaedt, J., Kern, M., Kosinski, M., Stillwell, D., Ungar, L., and Seligman, M. (2014). Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108.

Pennebaker, J. and King, L. (1999). Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77:1296–312.

Pennebaker, J. W., Mehl, M. R., and Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1):547–577. PMID: 12185209.

Pereira, A. C. (2021). Otimização do método disc de seleção de pessoas baseada em algoritmos genéticos e naïve bayes: Um estudo de caso em empresa do "sistema s" do paraná. 2:234–251.

Raguvir, S. and Babu, S. (2020). Enhance employee productivity using talent analytics and visualization. In *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, pages 1–5.

Santos, V., Paraboni, I., and Silva, B. (2017). Big five personality recognition from multiple text genres. pages 29–37.

Smith, B. L., Brown, B. L., Strong, W. J., and Rencher, A. C. (1975). Effects of speech rate on personality perception. *Language and Speech*, 18(2):145–152. PMID: 1195957.

Sumner, C., Byers, A., Boochever, R., Sumner, C., Byers, A., Boochever, R., and Park, G. (2012). Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. *Proceedings - 2012 11th International Conference on Machine Learning and Applications, ICMLA 2012*, 2.

Vu, X.-S., Flekova, L., Jiang, L., and Gurevych, I. (2017). Lexical-semantic resources: yet powerful resources for automatic personality classification.