PONTIFICAL CATHOLIC UNIVERSITY OF MINAS GERAIS

Graduate Program in Informatics

Richard Vinícius Rezende Mariano

# Exploiting curriculum vitae text features on employee classification using behavior archetypes

Belo Horizonte

2024

Richard Vinícius Rezende Mariano

# Exploiting curriculum vitae text features on employee classification using behavior archetypes

Research project presented at the Graduate Program in Informatics at the Pontifical Catholic University of Minas Gerais, as a partial requirement for obtaining the Master's degree in Informatics.

Advisor: Prof. Dr. Wladmir Cardoso Brandão

Belo Horizonte

2024

FICHA CATALOGRÁFICA
Elaborada pela Biblioteca da Pontifícia Universidade Católica de Minas Gerais

Mariano, Richard Vinícius Rezende
M333e Exploiting curriculum vitae text features on employee classification using behavior archetypes / Richard Vinícius Rezende Mariano. Belo Horizonte, 2024.
53 f. : il.

Orientador: Wladmir Cardoso Brandão
Dissertação (Mestrado) - Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática

1. Linguagem de programação (Computadores). 2. Processamento de textos (Computação). 3. Processamento de linguagem natural (Computação). 4. Aprendizado do computador. 5. Pessoal - Recrutamento. 6. Administração de pessoal. 7. Curriculum vitae - Análise do discurso. 8. Arquétipo (Psicologia). I. Brandão, Wladmir Cardoso. II. Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática. III. Título.



CDU: 681.3.091

Ficha catalográfica elaborada por Daniela Luzia da Silva Gomes - CRB 6/2505

Richard Vinícius Rezende Mariano

# Exploiting curriculum vitae text features on employee classification using behavior archetypes

Research project presented at the Graduate Program in Informatics at the Pontifical Catholic University of Minas Gerais, as a partial requirement for obtaining the Master's degree in Informatics.

---

Prof. Dr. Wladmir Cardoso Brandão –
PUC Minas

---

Prof. Dr. Zenilton Kleber Gonçalves do
Patrocínio Junior

---

Prof. Dr. Thierson Couto Rosa

Belo Horizonte, February 21, 2024.

**Acknowledgments**

Thank the Pontifícia Universidade Católica de Minas Gerais – PUC-Minas, the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES and the Sólides S.A. For the financial, knowledge and resources support.

## ABSTRACT

In the context of an organizational environment, the identification of an archetype can help place the right employees in the right companies. 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. The need to identify the archetype in the business environment arises to 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 their 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. To achieve this, this work explored the writing style of individuals from each archetype, and different classification approaches. Reaching an accuracy close to 65% in our binary classifier.



Keywords: People Analytics, Text Classification, Behavioral Classification, Natural Language Processing, Machine Learning, Transformers, Text Representation, Word Embeddings

## RESUMO

No contexto de um ambiente organizacional, a identificação de um arquétipo pode ajudar a colocar os colaboradores certos nas empresas certas. As organizações oferecem cada vez mais recursos para melhorar o desempenho, minimizar custos e alcançar melhores resultados. Uma organização são os indivíduos que nela trabalham ou prestam serviços. Portanto, o bom desempenho organizacional resulta diretamente do bom trabalho dos seus colaboradores. Surge a necessidade de identificar o arquétipo no ambiente empresarial para unir indivíduos com empresas, o que pode melhorar o ambiente organizacional e potencializar o desenvolvimento do indivíduo. Uma pessoa deixa vestígios de seu comportamento naquilo que produz, como por exemplo, seus textos. Alguns estudos apontam para a possibilidade de identificar um perfil comportamental a partir de uma produção textual. Neste trabalho buscamos identificar o arquétipo dos indivíduos dentro do ambiente empresarial a partir de seus textos curriculares. Para isso, este trabalho explorou o estilo de escrita dos indivíduos de cada arquétipo e diferentes abordagens de classificação. Atingindo uma acurácia próxima de 65% em nosso classificador binário.



Palavras-chave: Analise de Pessoas, Classificação de Texto, Classificação Comportamental, Processamento de Linguagem Natural, Aprendizado de Máquina, Transformadores, Representação de Texto, Word Embeddings

## LIST OF FIGURES

## LIST OF TABLES

## LIST OF ABBREVIATIONS AND ACRONYMS

**AI** *Artificial Intelligence*

**BoW** *Bag of Words*

**CNN** *Convolucional Neural Network*

**DA** *Data Analytics*

**DL** *Deep Learning*

**GPT** *Generative Pre-Trained Transformer*

**HR** *Human Resource*

**LIME** *Local Interpretable Model-agnostic Explanations*

**LLM** Large Language Model

**LSTM** *Long Short-Term Memory*

**MAE** *Mean Absolut Error*

**MT** *Machine Translation*

**NER** *Named Entity Recognition*

**NLP** *Natural Language Processing*

**PA** *People Analytics*

**QA** *Question Answering*

**RMSE** *Root Mean Squared Error*

**RNN** *Recurrent Neural Network*

**SVM** *Support Vector Machine*

**SVR** *Support Vector Regression*

**TC** *Text Classification*

**TL** *Transfer Learning*

**CONTENTS**

# 1 INTRODUCTION

An individual's set of behavioral patterns can define a behavioral profile. In psychology this is studied under the name of behavioral profiling or archetypes. Its focus is on identifying patterns of behavior, speech, and reactions when interacting with other people and the environment(SHAPIRO; BROWDER, 1990). From this information, groups of individuals with similarities are created. Each of these archetypes has skills that excel in different needs, demands, and environmental contexts. The identification of the individual's archetype is not just their categorization into groups, It is important to understand the archetype (GAL; JENSEN; STEIN, 2020). This task can significantly help individuals position themselves effectively and understand the underlying factors that drive their actions and reactions in various aspects of life, such as personal relationships, family dynamics, and professional environments.

In professional environment, Companies no longer rely solely on traditional recruitment methods; instead, they delve into the realms of behavioral psychology and advanced *Data Analytics* (DA) to identify the ideal candidates (ANGRAVE et al., 2016). A crucial objective of *Human Resource* (HR) teams is to determine a company's specific requirements and identify the individual profiles that best align with its unique context (BASSI, 2012). Furthermore, the use of technologies to assist in the recruitment task allows for a more egalitarian and inclusive process, minimizing human bias. Algorithms can assess candidates based on objective criteria, reducing the influence of subjective judgments on the selection process (VIEIRA et al., 2023).

Companies benefit from knowing the archetype of their employees as it improves significantly the management of their employees, allowing for the assembling of teams focused on a particular job, or combining skills of different employees to achieve great results (HEUVEL; BONDAROUK, 2017). In addition to helping the company know how to deal with employees facing difficulties or challenges (WABER, 2013). Knowing the archetype also makes it easier to get around problems, avoid unnecessary employee rotations and expand the development of the individual and consequently of the entire company.

Candidates are also benefited, as they are assisted through the process to become employees of companies that best fit their profile(Guoyin Jiang; Bin Hu; Youtian Wang, 2011). Avoiding companies that are not aligned with their needs or understandings. Besides, it becomes easier for them to enter organizations that genuinely resonate with

their values and aspirations. Choosing the right company unlocks numerous significant prospects for individual professional development, personal growth, and progress in one's career path (WABER, 2013). When workers discover the ideal match, they can thrive in an environment that develop their talents and skills, provides ample support, and promotes their overall well-being.

Several behavioral classification tools have emerged from studying psychological and behavioral profiles. They are focused mainly on the Eysenck Factors behavioral classifiers (EYSENCK; 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 automatically. Among other different models, *PACE* (VIEIRA et al., 2023) is available on the market as a Brazilian tool developed with a direct focus on the job market and based on the Brazilian business culture, to classify the archetype of individuals into four types.

Within this scope, this work proposes that an individual transmits their behavior and profile, that is, their archetype, in their texts. Psychology points out a correlation between personality traits and linguistic level, including acoustic parameters (SMITH et al., 1975), lexical category (PENNEBAKER; MEHL; NIEDERHOFFER, 2003), and several others. The *Natural Language Processing* (NLP) area is widely used in the study of texts, speeches, and discourses. This work seeks to explore whether and how each person leaves their mark, writing style, and personality in their textual production.

Today there are a few ways of determining an individual's archetype, such as BigFive (MCDOUGALL, 1932), DISC (MARSTON, 1928) and MBTI (BäCKSTRöM; BJöRKLUND; LARSSON, 2014). However, they are mostly very evasive methodologies and require time (SHAPIRO; BROWDER, 1990). It can cause discomfort, lack of interest and overthinking. Furthermore, current techniques, such as answering questionnaires or assessments with psychologists, cause concerns for individuals, who are often there looking for a job, or to improve their company. This may result in them not being completely honest when submitting to assessments. Our motivation is using automatic classification techniques in texts, to allow a faster, more efficient and less evasive process.

People Analysis is a recent topic, and it has been implemented in the job market more and more (TURSUNBAYEVA; LAURO; PAGLIARI, 2018). HR technological products are in hight demand (BASSI, 2012), due to the great growth in the use of information and technology to improve day-to-day business (ANGRAVE et al., 2016). Identifying an individual behavioral profile can help place the right employees in the right companies, and in the right positions (HEUVEL; BONDAROUK, 2017). With more information about its employees, a company can manage them better, knowing what motivates them, what are their main qualities and what their biggest challenges. Often the focus of the

company is not the same as the focus of the employee. When this focus is not the same, in addition to harming the effectiveness of the company, it prevents the individual from developing and growing professionally.

Since ancient times there has been a desire to identify and classify human behavior, and this still permeates in modern times. Knowing what type of personality a person has makes it easier to deal, mainly in business area (RAMLAWATI et al., 2021). There are several existing techniques to classify individuals into archetypes, but most of these methodologies are based on answering questions, questionnaires, psychologists' assessments, and other techniques that can be stressful for the individual, and not very efficient.

Therefore, an automatic assessment captures unstructured information, whether from videos, texts, or audio, and is capable of using this information to classify the individual's archetype automatically. Today, the information extracted from texts is used very little to help behavioral identification, but this work believes that the use of text, in addition to being innovative, can contribute to producing products focused on this market. The work then seeks to extract unstructured information from the texts, such as semantic issues, writing style, and transmitted emotions, and associate these unstructured data into classes of behavioral profiles, such as the archetype. With this problem in mind, this work raised the following hypotheses and Research Questions.

- Research Question 1: Does an individual transmit a behavior archetypes in his/her curriculum texts?

- Hypothesis 1: Individuals from different behavioral archetypes in an organizational environment, present different subjects in their curriculum texts.

- Research Question 2: Are techniques in NLP present in the literature, capable of extracting relevant information to be used in the identification of archetypes in the curriculum text?

- Hypothesis 2: Vector text transformation techniques, such as word embeddings and tokenization, combined with machine learning algorithms, like SVM and transformers, are efficient to identify and classify archetypes of an individual.

- Research Question 3: Are the models created in this work following the right path for archetype classification?

- Hypothesis 3: The decision-making of the classification models is consistent with the original definitions of the archetypes.

The main goal of this work is to have an effective model to identify archetypes automatically. The goal is to build a classification approach for predicting archetypes based on data extracted from curriculum text. In order to achieve the proposed goal, it is necessary to achieve the following specific goals: (i)Explore the characteristics necessary for the identification of an archetype; (ii)Construction of a dataset that aggregates texts made by individuals and their respective archetypes; (iii)Extract information from texts, and explore ways to better represent texts for processing; (iv)Finding the best algorithms for behavioral profile classification; (v) Make an automatic prediction model of archetypes, using textual information as a basis.

This work covers a whole area of study which is little explored in the literature, the use of texts with the objective of classifying a behavioral profile. By evaluating and proving the questions and hypotheses presented in this work, is possible to create classifiers that work automatically, capable of deducing archetypes from textual data information. In addition, this work offers a study on the approach of using Natural Language Processing, and on the relationship between texts and archetypes. Finally, this work presents new ways of exploring texts to extract information that can be used for various applications in addition to behavioral profiling.

This work is divided as follows. Chapter 2 defines important themes of the work. Chapter 3 presents the related works. The Experimental Setup is show in Chapter 4. The methodology is presented by Chapter 5. Chapter 6 displays the results obtained. A discussion of the results and the social impacts of the work is shown in Chapter 7. Finally, Chapter 8 brings the conclusion.

# 2 THEORETICAL BACKGROUND

This chapter addresses the main concepts necessary to understand this work, defining the main areas of study and techniques that will be used during development. First, we will explain what People Analytics is, a vital topic of our work. Next, we will explore some concepts and ways of categorizing existing archetypes, followed by an explanation of techniques and algorithms for classifying archetypes and texts.

## 2.1 People Analytics

*People Analytics* (PA) is a field aimed at guiding decisions concerning human resources and people management. This involves the utilization of data analysis and statistical techniques (GAL; JENSEN; STEIN, 2020), typically within a corporate context, to achieve enhanced optimization and deeper comprehension of the workforce. These insights encompass various facets, including but not limited to employee performance, engagement, recruitment, retention, and professional growth (TURSUNBAYEVA; LAURO; PAGLIARI, 2018).

This methodology has become increasingly present in HR teams embracing innovative technologies (GAUR; RIAZ, 2019; RAGUVIR; BABU, 2020) in their quest to enhance decision-making processes. The main focus is to identify information pertaining to behavior that can be used to monitor the performance, conduction, and outcomes. The applications of PA are multifaceted; in the business context, its primary goal is to enhance efficiency and productivity, minimize conflicts, and foster a more favorable work environment. For instance, by analyzing employee data, organizations can identify an employee's main skills, place them in the right task, identify their gap skills, and implement targeted training programs to improve their abilities.

To ensure the successful implementation of PA, it is essential to work with high-quality data sources (GAUR; RIAZ, 2019; GAL; JENSEN; STEIN, 2020). It is crucial to import data that contain as much information as possible, covering various aspects of behavior and performance. Leveraging advanced analytics methods like machine learning algorithms, predictive modeling, and natural language processing is necessary to extract valuable insights.

## 2.2 Behavioral Study

Human behavior is something of great fascination for humanity. People from different times, regions, and cultures sought a way to categorize, catalog, or divide people, their characteristics, and their behavior into different groups(VIEIRA et al., 2023). This grouping can be called a behavioral profile or archetype. By delving into the intricacies of individuals' behavioral profiles, we gain a deeper comprehension of their integration within society, discerning both the potential risks they pose and the invaluable contributions they can offer.

The number of behavioral profiles grouped throughout history varied mainly between four and five personalities. For example, the prophet Ezekiel(622-570 BC) made the division into four personalities, who saw humanity incarnated in four creatures: a lion, an ox, a man, and an eagle. The Greeks attributed human behavior to the four elements of nature: fire was moral, aesthetics and soul referred to water, intellectual to air, and physics to earth. Hippocrates(470-377 BC) considered the father of Western medicine, proposes that the human temperament is directly related to the balance of the four essential bodily fluids. Blood (happy temperament), black bile (somber temperament), yellow bile (enthusiastic temperament), and phlegm (calm temperament).

The relationship between human behavior with elements of nature and body parts is also found in Chinese culture. In Traditional Chinese Medicine (EIGENSCHINK et al., 2020), people are divided into five personalities, each one related to an element of traditional Chinese medicine and having an organ of the body that represents it, water (kidney), wood (liver), fire (heart), earth (pancreas) and metal (lung).

The physician and psychiatrist Carl Gustav Jung brings one of the most well-known classifications in modern times, dividing individuals into four groups: feeling, sensation, intuition, and thinking (JUNG; HULL, 1971). At the beginning of the 20th century, the American psychologist William Moulton Marston built the methodology DISC, which defines four main behavioral types: dominance, referring to control, power, and assertiveness; influence, related to communication and social relationships; stability, referring to patience and persistence; caution concerning organization and structure (MARSTON, 1928). Also, at the beginning of the 20th century, McDougall proposed the BigFive model, which defines five main factors influencing personality: neuroticism, extroversion, pleasantness, conscientiousness, and openness to experience (MCDOUGALL, 1932).

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

In recent years, models directly aimed at companies, with focus on improving the

work environment and hiring, have emerged. Among them, we have PACE (VIEIRA et al., 2023), a tool capable of providing more than 50 pieces of information about the individual, such as skills, pressure, energy and drive for work, motivational and withdrawal factors, ideal leadership style, and individual history as well as including the individual's archetype. To obtain this information, the individual must mark a series of adjectives, totaling 72, saying which of these adjectives he has and which he believes a company expects from him.

This work uses The PACE archetypes, Its methodology divides the profiles into 4 (Planner, Analyst, Communicator and Executor) 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 PACE approach allows for the possibility of various combinations and levels, which makes each personality unique.

**Figure 1 – PACE Report**


A brief explanation of each PACE archetype is described below.

- 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.

- 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 show a lot of determination, perseverance and willpower. 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.

## 2.3 Natural Language Processing

Natural Language Process is a subfield of *Artificial Intelligence* (AI) that centers on enabling computers to understand and process language as humans do. Studies in this field trace their origins back to the 1960s, with early attempts at natural language processing focusing primarily on rules and grammars. But it was in the 80s with Rule-Based Systems and heuristics to understand language that it really emerged(KHURANA et al., 2022). This marked a pivotal shift for NLP, as it transitioned from a mere exploration of handwritten rules to harnessing the potential of machine learning techniques (AHONEN-MYKA et al., 1998). In the 90s, statistical methods and support vector machines were introduced for specific NLP tasks (MANNING; SCHUTZE, 1999), as is the case with the Markov Model.

The renaissance of *Deep Learning* (DL), with *Recurrent Neural Network* (RNN) and *Convolucional Neural Network* (CNN), impacted several fields in the NLP area, producing improvements in performance, such as automatic translation tasks and entity recognition. But the great revolution came with the era of transformers. The BERT model, introducing a bidirectional approach to language pre-training, substantially improving performance across multiple tasks. Followed by GPT Models and others, which use transfer learning on pre-trained models on large data sets. A little about each of the models and algorithms is presented in the following Sections 2.4 and 2.5.

The field of NLP has several applications and studies. The *Named Entity Recognition* (NER) focuses on identifying and classifying entities, such as names of people, organizations, places and dates, in a text(NADEAU; SEKINE, 2007). Widely used for extracting key information from documents, sentiment analysis, and organizing large sets of textual data. An important task is Sentiment analysis, consisting of tasks such as evaluating the emotional tone present in texts, and classifying them as positive, negative or neutral (PANG; LEE, 2008). It can be used to evaluate customer feedback, monitor social media, analyze product reviews, among others.

*Machine Translation* (MT) is another existing application, where the main objective is automatically converting text from one language to another (BROWN et al., 1993). Being able to facilitate global communication, and assist in translating documents and online content. *Question Answering* (QA) is one of the most interesting applications, focused on answering questions based on information contained in texts (RAJPURKAR et al., 2016). They can be used by virtual assistants, automated customer support, and

improved search systems.

These applications before mentioned are just a few examples of the various possibilities that exist in NLP, as well as Text Representation and Text Classification, which will be explained more in depth in the following sections. However, there are several challenges along the way, such as ambiguities, linguistic and cultural varieties, in addition to the need for a quality data set (KHURANA et al., 2022). Despite this, the area has been developing, with several recent advances, and the future looks promising, especially with the use of AIs, such as chatbots. This research encompasses a fusion of some NLP tasks, with a particular emphasis on information extraction and text representation.

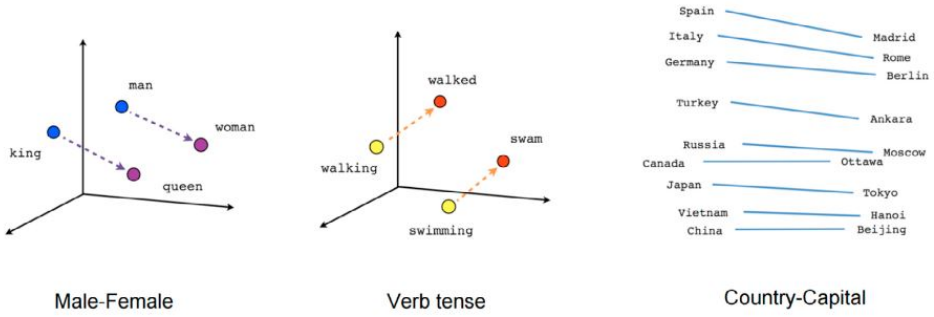
## 2.4 Text Representation

Texts are one of the main ways people use to communicate and convey their opinion. Whether through social media or even formal texts. Text representation is a subject of great interest in recent years, with the main approaches being based on machine learning. *Text Classification* (TC) techniques are becoming more and more advanced, due to the great demand in several fields of computing science. The way we are going to represent the texts is directly linked to the quality of the classification.

The representation in *Bag of Words* (BoW) was one of the first techniques for representing text in vector form. In this technique, each document is represented as a word frequency vector, where each position in the vector represents a word and its value represents the frequency of that word in the document. After that, natural language models began to emerge, one of the first word embeddings techniques is Word2vec (MIKOLOV et al., 2013), that consists in creating a high dimensional vector for each word. Similar to word2vec, we have GloVe (PENNINGTON; SOCHER; MANNING, 2014), a word counting model that generate dense word vectors. The GloVe considers global co-occurrence statistics while Word2vec considers only the context of each word. Still on the embeddings, the FastText (BOJANOWSKI et al., 2017) is an approach based on the skipgram where each word is represented as a bag of character n-grams. These three techniques were the most famous on the beginnings of word embedding study. The main difference between them is that fastText handles rare words better and is able to provide reasonable information even on words that were not seen in the training set. The word2vec and the Glove are simpler implementation and are more adequate for large corpus. In addition, the word2vec is able to fit many different tasks. In a word embedding similar words tend to have the same vector values and are grouped in the same block, as we can see in the example of a Word2vec representation in Figure 2.

Vaswani et al. (2017) introduced the Transformers, a new architecture for text

**Figure 2 – Word2Vec representation (JATNIKA; BIJAKSANA; ARDIYANTI, 2019)**




representation. This approach do not use recurrence and convolutions and is based only on attention mechanisms. The transformer is considered the state-of-the-art on text representation, they capture the meaning of a word in a specific context using attention mechanisms. Examples of transformers are BERT-like and GPT-like. The BERT (Bidirectional Encoder Representations from Transformers) is most popular transformer-based language (DEVLIN et al., 2018). The model processes text in both forward and backward directions, and proves to be more efficient than other models, because it considers the dependence between words. BERT and its variations have been extremely effective in a wide range of natural language processing tasks. It can be used in several tasks, including creating embeddings, pre-training and classification. Its success and efficiency has inspired the development of numerous transformer-based models for various language-related applications.

*Generative Pre-Trained Transformer* (GPT) (RADFORD; NARASIMHAN, 2018) is also a Transformer-based architecture and training procedure for natural language processing tasks. Training follows a two-stage procedure. First, a language modeling objective is used on the unlabeled data to learn the initial parameters of a neural network model. Subsequently, these parameters are adapted to a target task using the corresponding supervised objective. It became extremely popular with the emergence of automatic response chats, such as chatGPT. The most recent versions of GPT are among the most efficient in performing tasks with text, and demonstrated remarkable performance in tasks such as language generation, translation, summarization, and question-answering. It is extremely versatile in the application of NLP, just like BERT, and can also be used to create embeddings and fine-tuning.

## 2.5 Text Classification

Classification in general is fundamental in the field of machine learning and involves assigning labels or categories to objects or instances based on their observed characteristics or attributes (ALPAYDIN, 2020). Classification is a necessary step for algorithms to work with texts in TC approaches. In the beginning, text classification relied on manual rules defined by experts (SALTON, 1988), but this was a limited approach, and required specialized knowledge to create effective rules. It was with the emergence of word vectors, such as word2vec and GloVe, that the area began to take shape and bring better results.

With the emergence of neural networks, new approaches for text classification have been proposed. The RNN and CNN (KIM, 2014) techniques were successfully implemented, mainly long-memory RNNs, such as *Long Short-Term Memory* (LSTM) networks (HOCHREITER; SCHMIDHUBER, 1997), capable of capturing long-term dependencies. Recently with the introduction of Transformers, new possibilities for text classification have been enabled (DEVLIN et al., 2018). Transformers are capable of capturing bidirectional contextual relationships in sequences, significantly improving performance in text classification tasks. Still in this context, the task of *Transfer Learning* (TL) becomes relevant (PETERS et al., 2018), allowing the possibility of creating pre-trained models with large sets of data before adjusting them for specific tasks, generating more robust models with better responses.

One of the most basic algorithms used in several works focused on text classification is *Support Vector Machine* (SVM). The SVM are a set of supervised learning methods used for classification, and outliers detection. This algorithm is effective in high dimensional spaces (CORTES; VAPNIK, 1995). It can be used for both linear and non-linear problems. And although the focus of the algorithm is on supervised problems, it is possible to use SVM for clustering, through unsupervised techniques. The *Support Vector Regression* (SVR) is the variation of the algorithm, aimed at regression tasks.

Algorithms that are not commonly used for text classification can be used for specific tasks, such as application to data extracted from texts or pre-processing. This is the case with XGBoost, a decision tree-based machine learning algorithm that uses a Gradient boosting structure (CHEN; GUESTRIN, 2016). XGBoost Is a sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning. It is described as an algorithm able to solve world scale problems using a minimal amount of resources.

Within the task of text classification, it may be important to apply clustering techniques to organize text documents into groups or clusters based on their similarities. This task, called clustering documents, on an unsupervised execution, without the

need for prior labels, is an example of topic modeling techniques. Topic modeling stands as an unsupervised technique within the realm of machine learning, designed to uncover themes within document collections (CHURCHILL; SINGH, 2022). This method revolves around the generation of topics, which are essentially sets of words summarizing the content found in the documents. By doing so, it mitigates the requirement for extensive data or direct supervision, enabling the identification of significant topics based on extracted patterns from the corpus. One notable approach in topic modeling is BERTopic (GROOTENDORST, 2022).

BERTopic uses a pre-trained model to generate vector representations that encapsulate document information. To address the challenge of high dimensionality in these embeddings, dimensionality reduction techniques are applied. Subsequently, the model clusters these refined representations, where documents conveying similar meanings form clusters, each representing a distinct theme.

Machine learning models used in classification are often "black box", and we don't understand the reasons behind the predictions. To solve these problems, some tools exist, such as *Local Interpretable Model-agnostic Explanations* (LIME). The LIME is a tool focuses on helping to understand machine learning algorithms, and to know the reasons behind the predictions. In some more critical areas, understanding why a model makes a specific decision is crucial. And many times the construction of a model with better performance is left aside, due to the high complexity and difficulty in explaining.

In the case of LIME, the tool employs local surrogate models (RIBEIRO; SINGH; GUESTRIN, 2016). The objective of this model is to approximate the results of the black box models, however, focused on local training, not global training. The global training focuses on explaining the decision of the model as a whole, while local training focuses on being able to explain individual predictions.

# 3 RELATED WORK

This chapter presents the papers related with this work and those who are also working with classification or identification of behavioral profiles using text as a basis. A table with the main works it is also present, showing the behavioral model they are evaluating, features used, forms of evaluation, among other information.

Some studies are starting to use information obtained from text to identify and classify behavior. One of the pioneers in this study were Pennebaker et al. (PENNEBAKER; KING, 1999). They proposed an information extraction tool called LIWC. The LIWC is widely used in several works (GILL; OBERLANDER, 2002; MAIRESSE et al., 2007; GILL; NOWSON; OBERLANDER, 2009; GOLBECK et al., 2011; FARNADI et al., 2021; SANTOS; PARABONI; SILVA, 2017), some using only this tool, and others combining with other features.

The use of feature extraction is one of the main strategies for classifying behavior using text. There are relevant works in this regard, for example, MRC (MAIRESSE et al., 2007; GOLBECK et al., 2011) e MBSP (LUYCKX; DAELEMANS, 2008). But mainly several works explored different ways of extracting characteristics from the text, (ARGAMON et al., 2005; GOLBECK et al., 2011; PARK et al., 2015; MAJUMDER et al., 2017; SANTOS; PARABONI; SILVA, 2017; VU et al., 2017; ZHU et al., 2022), such as, grammar parser, lexical information, sentiment analysis, and Pos-tag. Works that use datasets taken from social networks also use information extracted directly from them (GOLBECK et al., 2011).

One of the main features used to explore and classify texts is the use of n-grams. This technique has been used for many years, as it provides good context for classification. Oberlander and Nowson (OBERLANDER; NOWSON, 2006; NOWSON; OBERLANDER, 2007), for example, used n-gram techniques, specifically bigrams and trigrams, in their work to measure accuracy in both binary and multi–class classification. Other works also make use of n-grams (LUYCKX; DAELEMANS, 2008; PARK et al., 2015), but this technique lost some of its strength with the emergence of word vectors.

The use of word vectors brought a new perspective on text classification, the words used, the order, frequency, and context become much more important when representing and classifying. This technique is a lot more present in current works (MAJUMDER et al., 2017; SANTOS; PARABONI, 2019; ZHAO et al., 2020; KARANATSIOU et al., 2022; JOHNSON; MURTY, 2023). Some of these representations are of simpler vectors,

as in the case of Mamjumder et al. (MAJUMDER et al., 2017), who uses word2vec. To more complex representations such as word embeddings used by Johson and Murty (JOHNSON; MURTY, 2023), the authors use the latest state-of-the-art techniques for representation, using transformers, in this case BERT.

What we can see is that the vast majority of works focused on behavior profile classification using BigFive profiles. And studies show that it is possible to extract relevant information of personality through written language analyzed (MORENO et al., 2021). Although some works do not explore all five profiles, but only some of them, to facilitate their analysis. However, we do have some exceptions, such as Gill and Oberlander(GILL; OBERLANDER, 2002) that exploit the Eysenck Factors and Luyckx and Daelemans (LUYCKX; DAELEMANS, 2008) that drive research in the MBTI typology. Works aimed directly at classifying behavior for the purpose of job applications are rarer. Jafari and Far (JAFARI; FAR, 2022) explore multiple proposed methods to predict behavior, analyzing the different approaches and metrics used.

All this description can be seen in the Table 1, together with various other information regarding each work cited in this section. Naturally, the articles mainly seek to measure accuracy and f1–score, with some seeking to explore more about the correlation of textual elements with the profile. Although the papers work with 3 or more profiles, the multi–class approach is little explored, and most works choose a binary approach, separating the profiles. This allows a better individual analysis of each of the profiles, prevents biases, and improves results. Finally, we can see that articles aimed at classifying behavior profiles based on text do not use explainability techniques, focusing only on visualization techniques.

**Table 1 – Papers that classified behavioral profile from textual productions**

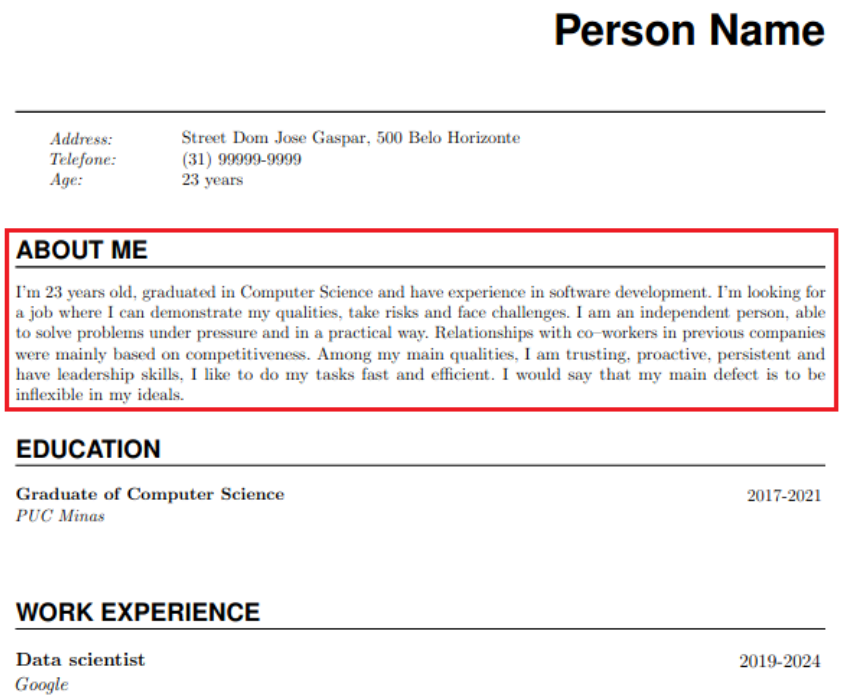
| Ref | Description | Behavior model | Features | Evaluation | Measure | Explainability |
|---|---|---|---|---|---|---|
| Pennebaker and King (1999) | Explore the impact of linguistic style in the exploration of personality | BigFive | Feature Extraction LIWC | - | Correlation | no |
| Gill and Oberlander (2002) | Study the impact of a profile in people's language production | Eysenck Factors | LIWC | Rank | Correlation | no |
| Argamon et al. (2005) | Prediction of Personality using Lexical features | BigFive | Grammar Parser Lexical | Binary | Accuracy | no |
| Oberlander and Nowson (2006) | Using personal blogs for create a automatic classification of authors personality | BigFive | n-grams | Binary Multiclass | Accuracy | no |
| Nowson and Oberlander (2007) | Create a automatic classification of authors personality with a large number of blogs | BigFive | n-grams | Binary Multiclass | Accuracy | no |
| Mairesse et al. (2007) | Recognition of Personality using both Conversation and Text | BigFive | LIWC MRC | Binary Rank Regression | Accuracy Loss | no |
| Luyckx and Daelemans(2008) | Present a corpus and personality prediction | MBTI typology | MBSP n-grams lexical | Binary | Accuracy F-score | no |
| Gill et al. (2009) | Explore the use of blogs from each personality | BigFive | LIWC | - | Standard deviations | no |
| Golbeck et al. (2011) | Predict users Personality using public information on twitter | BigFive | Twitter information Sentiment analysis LIWC, MRC | Rank Regression | Correlation Loss | no |
| Farnadi et al. (2013) | Inferring a user's personality traits from Facebook status updates | BigFive | LIWC Feature Extraction | Binary | Precision Recall F-score Correlation | no |
| Park et al. (2015) | Assessing personality using an open-vocabulary analysis of Social media | BigFive | Feature Extraction N-grams | Binary | Correlations | no |
| Majumder et al. (2017) | language from social media. | BigFive | Feature Extraction word2vec | Binary | Accuracy | no |
| Santos et al. (2017) | Different text genres using for recognition of personality | BigFive | LIWC Lexicon | Binary | F-score | no |
| Vu et al. (2017) | Multiples datasets with multiples features for identify the best combination | BigFive | Lexical (Wordnet) Sentiment Analysis POS-tag | Binary Rank | F-score | no |
| Santos and Paraboni (2019) | Apply different word-vectors representation on text from facebook | BigFive | Word-vectors | Binary | F-score | no |
| Zhao et al. (2020) | Predict the personality of social network users | BigFive | Word-vectors Attention Information | Binary | F-score Precision Recall | no |
| Karanatsiou et al (2022) | Identify correlation between personality and relational traits | BigFive | Word-vectores | Cluster Regression | Correlation Loss Accuracy | no |
| Johnson and Murty (2023) | personality detection using a new enhanced psycholinguistic knowledge graph-based | BigFive | Emotion Lexicon Word embeddings | Binary | Accuracy F-score | no |

# 4 EXPERIMENTAL SETUP

## 4.1 Dataset

The dataset of this work is composed of curriculum texts in Portuguese, produced by people, describing themselves. This information is easy to obtain, and is available in almost all CVs. There is a common example of a CV in Figure 3. Containing personal information, a summary about the person, academic education and work experiences. The information use in this work is just the "about me" session, marked in red.

**Figure 3 – Curriculum vitae example**




The dataset of curriculum texts, used in this work is a private data base, extracted by the company responsible for the PACE tool, and it is in *csv* format. This dataset consists of 26,636 instances. Each instance consists of a text written by an individual, the respective percentages referring to each archetype of that individual, where A means Analyst, C refers to Communicator, E to Executor, and finally, P means Planner. In the last column there is the formation of their final archetype combination, as explained in Section 2.2. An example instance is shown in Figure 4.

The classes composition of the dataset is divided as follows: 38.5% have the Analyst archetype above or equal to 25%. 50.81% have the Communicator archetype, 58.67%

**Figure 4 – Instance example**

| About me | P | A | C | E | Archetypes |
|---|---|---|---|---|---|
| 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 | 23.16 | 16.48 | 29.40 | 30.96 | EC |

Executor, and finally, 51.96% have the Planner archetype. Remembering that each individual can have 1 to 3 archetypes, with percentages greater than or equal to 25%. The complete composition of the dataset following the number of instances for each possible combination can be seen in Table 2.

**Table 2 – Composition of dataset**

| Main Planner | | Main Analyist | | Main Communicator | | Main Executor | |
|---|---|---|---|---|---|---|---|
| P | 882 | A | 478 | C | 974 | E | 1671 |
| PA | 2731 | AC | 245 | CA | 297 | EA | 1004 |
| PC | 1165 | AE | 686 | CE | 3274 | EC | 3963 |
| PE | 876 | AP | 2461 | CP | 1247 | EP | 1122 |
| PAC | 164 | ACE | 30 | CAE | 27 | EAC | 63 |
| PAE | 209 | ACP | 75 | CAP | 53 | EAP | 171 |
| PCA | 139 | AEC | 37 | CEA | 46 | ECA | 91 |
| PCE | 162 | AEP | 243 | CEP | 326 | ECP | 299 |
| PEA | 188 | APC | 120 | CPA | 93 | EPA | 185 |
| PEC | 114 | APE | 300 | CPE | 238 | EPC | 182 |
| total | 6630 | total | 4675 | total | 6575 | total | 8751 |

## 4.2 Features

This work carried out multiple experiments, and each one requires a different type of feature, or way of manipulating the database. However, two main types of features are addressed, the use of text representation through vector and extraction of characteristics. It is noticeable in the literature that the main results obtained come from the combination of extracted characteristics with a word vector. Experiments will be carried out with both types of features, to evaluate whether these features are also efficient for the dataset of this work. And if so, these features will be combined to generate a better model.

### *4.2.1 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 techniques, such as TF-IDF, to more complex techniques like word embeddings. These word representation techniques play a crucial role in capturing the semantic meaning and context of words, enabling the models to understand and process textual data more effectively.

For text representation, two distinct approaches were chosen. Firstly, the tokenization technique from Keras, provided by Tensor Flow. A simpler technique, which does not require prior learning, but which proves to be very efficient in representing text. In this method, each word has its numerical representation, ultimately generating a vector of numbers. Is necessary apply a pad sequence to keep the vectors the same size, the final size defined for each instance was 250. In this case, only those 250 most frequent words within the complete vocabulary remain in each instance. Texts with less than 250 words are completed with numeric data "0".

Additionally, the use of word embeddings, specifically the BERT and GPT models, for representation in this study. These choices allow us to explore and compare the effectiveness of both methods in capturing the underlying semantics and contextual information within the text data. All embeddings in this work were adjusted to a size of 256. In the case of BERT, two models were used, one multilingual uncased, and one in English also uncased. While the GPT embedding used was the "ada" model from opneAI.

In the experiments that used text vectors, pre-processing techniques were necessary for a clean execution. The data was sanitized to facilitate representation and classification learning. It is import to be very careful with the pre-processing because it 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 techniques used in this work are:

- Remove special characters, punctuation and accentuation;
- Remove stopwords, the most common words in a language;
- Transforming the whole text to lowercase letters;
- Lemmatization. Grouping the inflected forms of a word so that they can be analyzed as a single item.

### *4.2.2 Characteristics Extraction*

A greater exploration of what is being said in a text is possible through the extraction of characteristics. The idea is to go beyond just the text itself and obtain information about its composition. For this, a pos-tagger tool was used as an aid. This tool extracted from the text the number of times it has each grammatical class.

It may be necessary to flatten the data, considering that there are texts of different sizes. For this, the size of the texts was extracted, obtaining the number of total words. This also allow for obtaining the proportion of each grammatical class in relation to the total number of words, which is basically having a percentage of representation of that grammatical class in the texts.

The Pos-tagger tool used in this work is open-source and available on Github*. 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 13 features.

Characteristics extraction is widely used in literature. It is a more explanatory and intuitive approach compared to text-vector. The idea of using Characteristics extraction is to combine it with the text vector, to build a more robust model, with more information.

*github.com/inoueMashuu/POS-tagger-portuguese-nltk

# 5 METHODOLOGY

This work aims to extract information, from a dataset of CV texts, aim towards classifying people in archetypes. The archetypes are the classes of this work, with a total of four classes, one referring to each of the archetypes: Planner, Analyst, Communicator and Executor. The order of experiments that will be carried out can be seen in the figure 5. While the classification architecture of the experiments is shown in figure 6 Following is an explanation on obtaining the dataset and going through each of the necessary steps to reach the final objective, which is to have efficient classifiers for archetypes.

**Figure 5 – Methodology path**

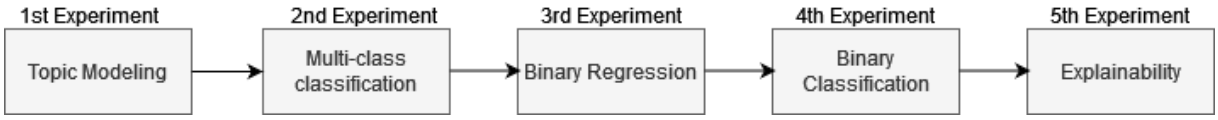
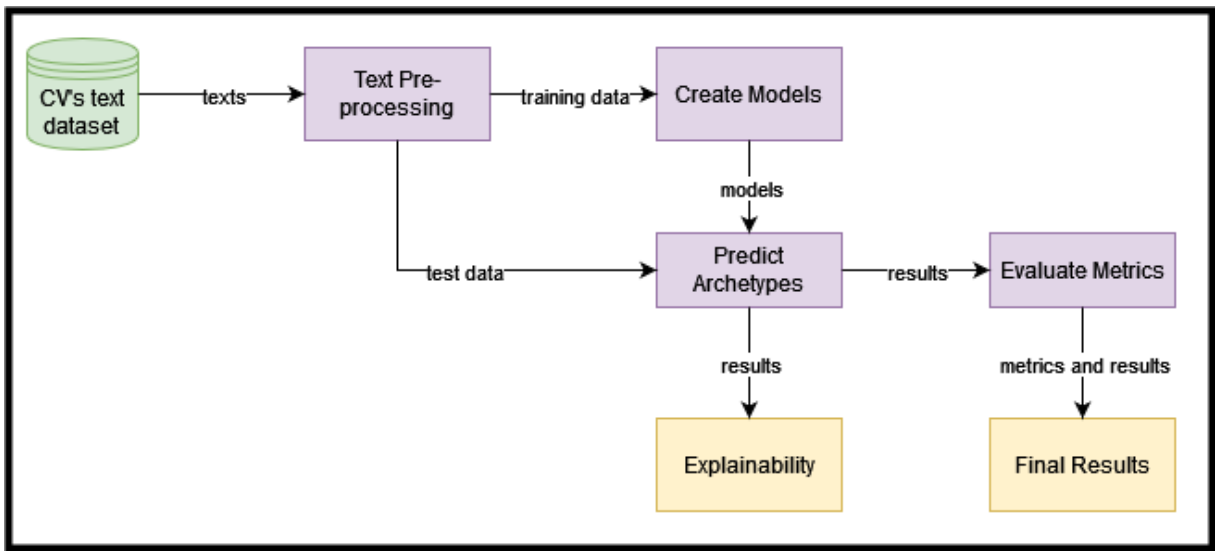



**Figure 6 – Architecture of classification approach**




To carry out the proposed experiments, the CV dataset is necessary, it contains the curriculum texts and the distribution of the percentage of archetypes. The curriculum texts is the information use to create and train the models. The archetypes are the classes, having a total of 4 classes, since we have 4 archetypes. With this dataset, we will select the features and what pre-processing technique to use. Initially, we chose to use two types of features, characteristics extraction and textual representation. This second

one ended up being more used, due to better performance in the experiments carried out afterwards. The application of pre-processing mainly consists of removing special characters, punctuation, accents and stop-words. Furthermore, lemmatization techniques were applied to group the inflected forms of the words.

The first experiment aims to answer question 1, which is that there is evidence that points to differences in the texts of individuals of different archetypes. A topic modeling study is carried out, to help understanding the main topics covered in each archetype group and which topics are the most relevant in each of the classes. This experiment presents the first results of the work, and contributes to decision-making for the next experiments.

The following experiments require dividing the dataset into training and testing. A division of 80% training and 20% testing was defined after pre-processing the data. The training portion will be submitted to the models according to each demand and the test portion will be used later to check the accuracy of these models. The forms of textual representation used in this work were Keras tokenization, BERT and GPT embeddings.

The main objective of the following experiments is to predict the individual's archetype solely based on their written texts, for this we use the training data to create the models. These experiments aim to answer question 2. The archetype identification approaches in this work are binary regression, multi-class classification and binary classification. The latter having a greater variety of explorations. For the second experiment, the multi–class approach, the SVM algorithm was used, with its own evaluation metrics, starting from more simplistic analysis to an analysis that can better explain the decisions. Each instance can have more than one class, therefore, a multi class approach is relevant. Allowing for evaluating the capacity of prediction models for more than one class at the same time. Token representation using Keras tokenization was also used in this approach. The choice to use the SVM algorithm is justified by the ease of the algorithm, which requires less resources and time. In addition to the high number of works in the literature that use it. It is interesting to use it to create a baseline of results. Considering that the classes in this work have never been previously used for automatic classifications.

The third experiment, is binary regression approach tried to directly predict the profile percentage, using the SVR algorithm and calculating the error of the models. Calculating the error gives an average of the distance from the original values to the predicted values. This approach is very relevant for predictions of continuous values. It is interesting to do this for this problem, considering that the intensity of the classes of instances are not fixed. Evaluating this error is interesting, because two individuals with the same archetype composition can have completely different percentages for those archetypes. In binary regression, Keras tokenization was used as a feature.

The binary classification approach is the four experiment. Has become the most exploratory in this work, creating a different classifier for each of the classes. The choice for this approach is due to the ease of classifying and explaining the models, and for its vast literature. Different models were created, combining some features and algorithms. Using the features of characteristics extraction, two algorithms were applied, SVM and XGBoost. SVM, for the same reason as used in previous approaches, and XGBoost, a supervised algorithm capable of excellent results using numerical data.

The textual representation features were also applied to binary classification. These representations in this approach were more diverse. Keras tokenization was used, as well as word embedding. Initially, two BERT embeddings were used, one multilingual, as the texts are in Portuguese, and one in English, when translating the text, since embedding in English is more robust. To perform the features, the SVM algorithm was used alongside a neural network called "BERT for classification", a neural network built directly to be used in BERT-type embedding. The last step was a final binary approach, combining GPT-type embedding with fine-tuning provided by the openAI platform. This approach has been considered state of the art in text classifications, mainly with the popularization of tools like chat-gpt.

After several experiments and creation of multiple models, the test data was submitted to these models to be able to evaluate their results, according to each previously decided metric, such as error, accuracy, and others depending on each approach. These results are responsible for answering the questions regarding the ability to classify archetypes using curriculum text, more specifically, texts in the dataset. To add an additional layer of discussion regarding the results, explainability was applied to the model that performed best, which helps to show how that model took certain decisions in a text sample, this is the fifth experiment, focusing on answering question 3. The final idea is to take these models and create a tool or API that can receive CV texts and output the archetypes of the individual who wrote it.

# 6 RESULTS

This chapter presents the results of this work. First a topic modeling study, exploring the themes covered in the dataset. Then, regression experiments carried out in a binary way, and classification experiments carried out in a multi–class and binary way. Finally there is an experiment with the objective of predicting the adjectives of each archetype, and also an explainability for the best model.

## 6.1 Topic Modeling

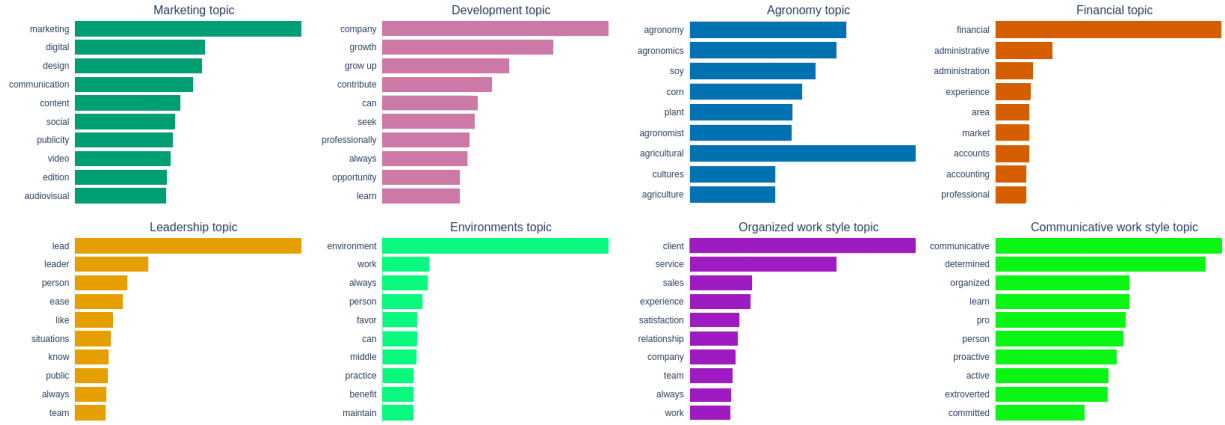
Employing the topic modeling technique, specifically using BERTopic, allows identifying word patterns that capture common themes related to behavior profiles. Initially, the main topics found in general CVs were extracted. After the extraction, an analysis on these topics concerning different behavior profiles was performed. The topic modeling method is interpretable, allowing to understand how each behavior profile is expressed in text.

Figure 7 represents a characterization of the dataset using the topic modeling technique. Eight main subjects addressed in curriculum texts were chosen to be presented. They include topics related to: (i) marketing; (ii) development; (iii) agronomy; (iv) finance; (v) leadership; (vi) environments; (vii) organized work style; and (viii) communicative work style. Topics are arranged in order of relevance. The words in the bars for each topic are the most frequent words that represent that topic. For example, in the topic of Marketing, the most common word is the word ‘´´marketing” itself, followed by the word ´´digital”, and so on.

The next step, was extracting the main topics from each of the archetypes. By extracting the topics contained in the texts of each profile, four main ones from each of the archetypes were selected. The view displays the main topic of the theme and then the subtopics, specifically, the words that belong to that main topic.

When analyzing general subjects in the CVs of people with the planner profile in Figure 8, it is suggested that the four predominant topics featured in their CVs are related to finance, marketing, development, and agronomy. When it comes to the planner archetype, we do not notice major differences in its topics in relation to the other archetypes, as will be presented below. This archetype does not have any single topic among its main ones, they are all repeated in some way in other archetypes. However,

**Figure 7 – All topics**

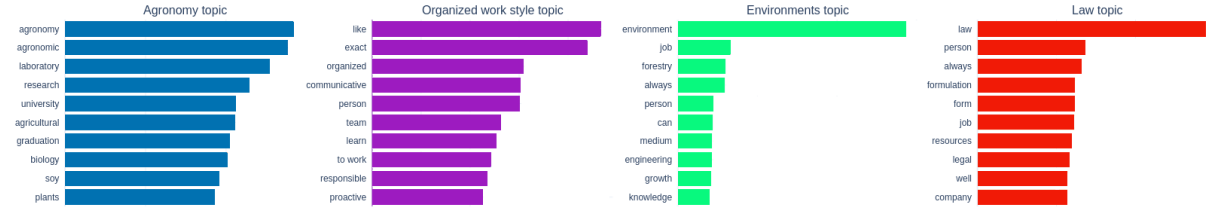



the order of relevance of these topics is not the same.

**Figure 8 – Planner topics**

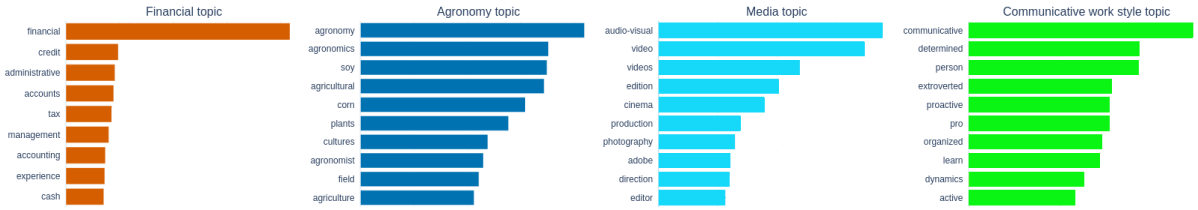

Regarding the analyst profile, see in Figure 9, the four predominant topics seem to be related to agronomy, organized work style, environments, and laws. For the work style, this profile tends to write about the characteristics of punctuality, organization, and responsibility.

**Figure 9 – Analyst topics**



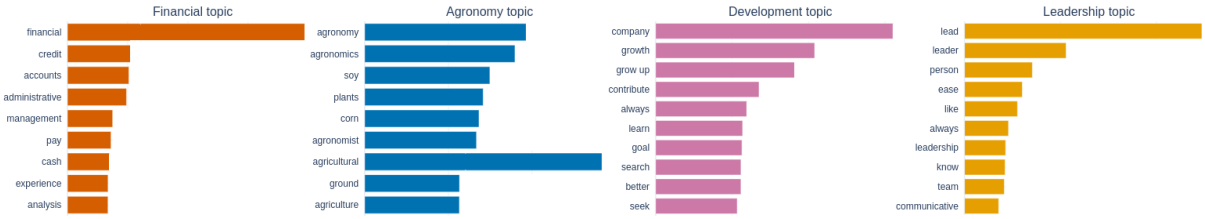

For the communicator profile, topics related to the financial sector and agronomy also appear. Communicators also seem to write about media, including audio visual, video, editing, cinema, and photography. This profile also tends to write about the communicative work style. Terms such as communicative, determined, extroverted, and proactive are frequent in the texts of the communicator profile. The main topics of the communicators are in the Figure 10.

**Figure 10 – Communicator topics**




Finally, for the executor profile, is present in Figure 11 the topics related to the financial sector, agronomy, and development predominate, in addition to topics related to leadership. This is a more solution-oriented profile, and who do well in more exact tasks, as is the case in the financial sector. They are more objective and have leadership characteristics. They have a lot of confidence and always seek to develop themselves. These characteristics can help explain the main topics covered by this archetype.

**Figure 11 – Executor topics**




It is possible to notice that some topics prevail in more than one archetype, but the relevance of the topics is different, as is the order of the subtopics. For example, agronomy is present in the four main topics of all archetypes. This may show a common characteristic not only of the people who make up the dataset, but also of the environment in which they belong, of a country that demands and depends a lot on the agronomic area for its economy. This characteristic can help explain the difficulty in classifying archetypes, as they are similar.

Experiments on topics between archetypes contribute to answer question 1. Considering the main topics of each archetype, It's noticeable that some archetypes have the exclusive presence of certain topics, and the level of relevance of some topics varies between each archetype. This partially helps proving hypothesis 1, but there is not enough evidence for validation, additional experiments are still required. Furthermore, there is a certain base bias for some specific topics, as is the case with the topic of agronomy.

## 6.2 Multi-class Classification

The first approach is viewing the problem as a multi-class problem. Each instance can have from 1 to 3 classes. The most common being having 2 of the 4 classes available, which occurs 72.21% of the time in the dataset, while 14.45% have only one class and 13.34% have 3 classes.

Although the instance are 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, the problem was divide into 4 scenarios of analysis, so that it is possible to observe different aspects of the behavior and performance of the classification.

*Analysis 1 (A1):* Hit only the Main Archetype. If the highest probability in the classifier output is equivalent to the Main Archetype of the instance, then there is a hit.

*Analysis 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.

*Analysis 3 (A3):* Hit some profile archetype. If any probability of the classifier equals 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.

*Analysis 4 (A4):* Each archetype is considered a hit or miss. This analysis is an extension of Analysis 3, but it 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 there is 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. This case also facilitates the calculation of the classifier error.

It is important to remember that the PACE tool that defines the individual's profile, uses the threshold of 25% to define the individual's archetypes, and therefore, this threshold was chosen in the experiments scenarios.

This multi–class classification approach allows for an initial overview in the analysis of the problem. The algorithm used was the SVM in a One vs One heuristic method. The results can be seen in Table 3. The A1 and A2 analysis are limited, since they consider 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 PACE.

**Table 3 – 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 bit of each archetype, even having more than one dominant profile.

### 6.3 Binary Regression

The multi-class approach proved to be difficult to perform and analyze, therefore, the alternative of dividing the problem into binary problems is an alternative. A binary approach allows us to divide the one big problem into four smaller problems. The first binary analysis used regression.

With regression algorithms it is possible to use continuous data, in training and prediction of data. This approach allows working directly with the percentages passed by the PACE. In that case, four different regressors were built, one for each archetype. These regressors will be made using the SVR algorithm.

In this case the metric that matters is the difference between the right answer and what was predicted, the error. Two techniques were chosen for error calculation, *Root Mean Squared Error* (RMSE) and *Mean Absolut Error* (MAE). Both metrics calculate the distance between actual values and predictions, but the RMSE squares this value for each instance before calculating the average, which ends up putting weights on the sum. This metric suffers from data where there are many outliers. The MAE calculates exactly the average of the distances between actual values and predictions. For both error metrics, the smaller the value, the better the model result.

The results of each error metric can be seen in Table 4, together with the accuracy obtained. To calculate accuracy, was follow the PACE standard of considering the threshold of 25%, if the prediction is above or equal to 25% it indicates that it has the archetype, if it does not, it means it does not have it.

The MAE metric has better results than the RMSE, which points to the existence of some outliers that generate an increase in the RMSE. Considering that profile percentages can vary from 0 to more than 50%, an error between 4 and 7 shows very positive

**Table 4 – Regression report**

| | RMSE | MAE | Accuracy |
|---|---|---|---|
| P | 5.98 | 4.49 | 0.52 |
| A | 7.02 | 5.24 | 0.62 |
| C | 7.06 | 5.36 | 0.50 |
| E | 7.05 | 5.24 | 0.58 |

results, however when looking at accuracy, the results are less satisfactory. Despite this, an accuracy of 62% was achieved in the Analyst archetype, and 58% in the Executor archetype.

### 6.4 Binary Classification

Still following a binary approach, predicting the percentage of the archetype showed to still be very complicated. To further simplify the problem, it was decided to follow a classification path.

In the binary classification, the problem was also divided 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; and (iii) allow for a better analysis of the decision making of the algorithms, which will allow a greater explainability of the models.

There are four classifiers, each focused on classifying one of the archetypes in the dataset. In this way, four datasets were generated, deriving 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. Data balancing was applied to each of these four datasets as needed. For example, there are more executors than non-executors, so the number of executors was decreased in the dataset. In the case of the Analyst, there are more non-Analysts than Analysts, so the number of non-Analysts was reduced. Accuracy and F1–score were chosen for evaluation metrics. Below is an explanation of each metric, and formulas based on Table 5.

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

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

**Table 5 – Confusion Matrix**

| | | Predict Valor | |
|---|---|---|---|
| | | Yes | No |
| Real | Yes | True Positive TP | False Negative FN |
| | No | False Postive FP | True Negative TN |

b. *F1–score:* Combines precision and recall in a balanced way. Considering the number of times your model hits in relation to the total number of times it tries to hit (precision). And also the number of times your model hits in relation to the total number of times it should have hit (recall).

$$Precision(P) = \frac{TP}{TP + FP},$$

$$Recall(R) = \frac{TP}{TP + FN},$$

$$F1 - Score = 2 * \frac{P * R}{P + R}$$

### *6.4.1 Characteristics Extraction*

In this experiment, only the features of characteristics extraction were used. There are a total of thirteen features, including 12 grammatical classes (parts of speech) and one word counting feature, as explained in the session 4.2.2. Extracting the grammatical classes of text is important to help identify whether there is a difference in the amount of words used in each of these classes according to the profile. A certain profile can use more adjectives than the others and less nouns, for example.

Two classification algorithms were chosen for this task, the first, the SVM algorithm, which was also used in previous tasks in this work, and in addition the XGBoost algorithm, to provide a comparison of results. The results of both classifications can be seen in the Table 6.

The results obtained from the extraction of characteristics were low in both algorithms. Despite this, this line of experimentation still has a future, using another more efficient characteristics extraction method.

**Table 6 – Binary classification report: characteristics**

| | SVM | | XGBoost | |
|---|---|---|---|---|
| | Accuracy | F1–score | Accuracy | F1–score |
| P | 0.52 | 0.50 | 0.52 | 0.52 |
| A | 0.51 | 0.51 | 0.51 | 0.51 |
| C | 0.53 | 0.53 | 0.51 | 0.51 |
| E | 0.54 | 0.52 | 0.52 | 0.52 |

### *6.4.2 Text-vector*

In this experiment, two algorithms were used, a more basic algorithm and widely used in previous works for text classification, the SVM. And a second algorithm, the BertForClassification (BFC), a neural network, developed to work with BERT-type embeddings. And to work with these algorithms, 3 forms of textual representation were proposed, tokenization, a textual representation made available by the tokenize library in Python. And two types of pre-trained BERT word Embeddings, one in Portuguese and the other in English.

The Accuracy and F1-score metrics were calculated, and the results obtained can be seen in Table 7. The first column refers to the results combining tokenization with SVM. The second column is the combination of the SVM algorithm with the use of BERT word-embeddings in Portuguese. Next, there is the combination of BERT word-embeddings but using the BFC. Finally, in fourth column followed the same setup as in the third column, but in this case, the text was translated into English before transforming it into embeddings. Taking into account that the embedding models in English are more robust than in other languages such as Portuguese, the original language of the CVs.

**Table 7 – Binary classification report: text-vector**

| | SVM + Tok. | | SVM + BERT | | BFC + BERT | | BFC + eng. BERT | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | F1-score | Acc. | F1-score | Acc. | F1-score | Acc. | F1-score |
| P | 0.65 | 0.62 | 0.50 | 0.51 | 0.59 | 0.59 | 0.60 | 0.60 |
| A | 0.63 | 0.63 | 0.51 | 0.51 | 0.57 | 0.57 | 0.60 | 0.59 |
| C | 0.63 | 0.60 | 0.50 | 0.50 | 0.59 | 0.59 | 0.59 | 0.59 |
| E | 0.63 | 0.63 | 0.50 | 0.50 | 0.57 | 0.57 | 0.60 | 0.60 |

In theory, using word-embeddigs with neural networks, are generally superior to other representation and classification techniques. In the case of the dataset of this work, the use of tokenization as a representation together with the SVM algorithm proved to achieve superior results.

### *6.4.3 Large Language Model*

In recent years, the popularity of artificial intelligence has exploded. The use of tools that help users obtain quick information, with simple questions, is increasingly present in everyday life. One of these tools that became very popular was Chat-GPT, provided by OpenAI*. This type of model is also called Large Language Model (LLM) (BROWN et al., 2020) and, like the OpenAI's GPT, consist of neural networks with millions or even billions of parameters. They are designed to understand and generate human-like text based on the patterns and structures it learns from vast amounts of data.

The use of models based on GPT transformers has become the state of the art in problem solving. And one of the ways to reach the goal is the use of fine-tuning (DODGE et al., 2020). This technique allows additional training of artificial intelligence, with custom data, helping the model adapt to the nuances and characteristics of the target, improving its performance and efficiency.

The first step of this approach was to take the training dataset, the same one used in previous experiments, and apply fine-tuning. The model chosen was "ada" from openAI, Although there are some more robust models, the *ada* model stands out for being efficient and having a low cost and requiring less time. Fine tuning was perfomed with the curriculum texts, each text and its respective class are passed. After fine-tuning is performed, the test dataset is passed, and for each text the model presents a response from the class. A binary approach was also used in this case, generating a total of four fine tuning models. The results can be seen in Table 8, they are compared with the best obtained so far.

**Table 8 – Binary classification report: fine-tuning openAI**

| | SVM + Tok. | | Fine-tuning openAI | |
|---|---|---|---|---|
| | Accuracy | F1-score | Accuracy | F1-score |
| P | 0.65 | 0.62 | 0.61 | 0.61 |
| A | 0.63 | 0.63 | 0.65 | 0.53 |
| C | 0.63 | 0.60 | 0.58 | 0.56 |
| E | 0.63 | 0.63 | 0.57 | 0.58 |

The use of GPT embeddings for fine tuning is one of the state of the art in textual classification, despite this, this approach was not able to surpass the best classification results of this work, which is the combination of SVM with tokenization representation.

*openai.com

## 6.5 Classification of Adjectives

A step back was taken and we tried a different approach, considering the difficulty of directly identifying the archetype. In PACE, as detailed previously, the candidate needs to answer a questionnaire with a series of adjectives, where they must mark those that they consider they have, and those that they believe a company demands from them. In this approach, we proposed to predict from the user's text whether or not they would mark a certain adjective. Thus creating a cluster of classifiers, one referent for each adjective, in a total of 144 classifiers.

Upon obtaining the result of these classifiers, they were submitted to a PACE API, simulating a candidate marking, and the corresponding archetype results obtained. Then, it is possible to compare these results, with the correct archetype, and evaluate the accuracy and F1–score. The results can be seen in the Table 9.

**Table 9 – Binary classification of adjectives**

| | Accuracy | F1-score |
|---|---|---|
| P | 0.48 | 0.19 |
| A | 0.58 | 0.17 |
| C | 0.49 | 0.55 |
| E | 0.63 | 0.77 |

This approach, despite bringing a different perspective in an attempt to improve the results, had no effect, the results were inferior to those obtained previously and we noticed a certain overfitting on the PACE API when generating the archetypes based on the predictions.

## 6.6 Explainability

The goal of explainability is to understand the reasons that lead a machine learning algorithm to take a certain decision. Machine learning algorithms tend to be, in general, a "black box". Usually, it just extracts certain metrics, without understanding the reasons behind the predictions. In some classifiers, such as decision trees, it is possible understand a little about the path taken by the algorithm. But in more complex cases, such as neural networks, the path is not clear, due to a large number of parameters, which can be thousands or even millions, understanding cannot be done quickly, which prevents quick decision-making.

This explainability technique can help us understand where the models are going wrong, and help make necessary adjustments and know when the models are right to understand if the path taken makes sense. In this work, explainability will help us answer

and validate hypothesis 3. Whether the models in this work are consistent with the archetype definitions.

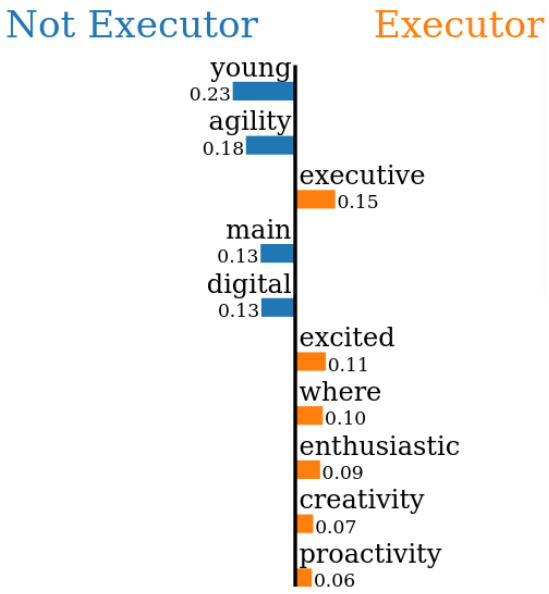
First, we searched in the PACE for the main words that describe each profile, these words can be seen in table 10. The idea is to associate these words with the words that the model judges to be relevant in the classification.

**Table 10 – 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. This analysis used the binary classifiers that performed best, SVM + tokenization. It also used the binary model of Executor archetype. First, analyzing the sample taken from the dataset, and then applying LIME explicability as can be see in Figure 12, the full text will not be displayed for privacy reasons.

**Figure 12 – Lime report from curriculum sample**

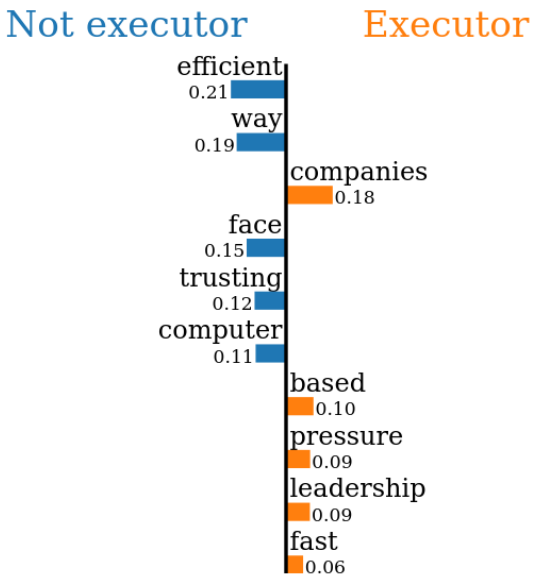


The explanation of the figure shows the local explainability of a curriculum text. On the right, It is the features that have a positive correlation with the output of the analyzed class, and on the left a negative correlation. For example, the word “executive” is the word that has the highest correlation with the class to be predicted. While the words “young” and “agility” are the main words with opposite correlation to the analyzed class.

Figure 13 shows the application of local explainability to a text created for research purposes. See the text bellow.

> 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.

**Figure 13 – Lime report from a self-made text**




Is shown in table 10, words like “active” and “leader” are used to describe the Executor archetype, as well as in figure 13 that words like “fast” and “leadership” has a positive correlation with the Executor class. However, some words like “efficient” were negatively correlated with the performer profile. These variations occur probably because each person is formed by the four archetypes.

This explainability experiment shows that the models use some words from the text to classify the archetype, which match what is used for the official definition of these

archetypes. However, some of the words used go against what was expected, which can be explained by the influence of other classes, since every individual has some of each archetype. Therefore, Hypothesis 3 cannot be affirmed, the models' decisions are not consistent with the original definitions, but more experiments are still needed, with more data and examples, exploring all archetypes.

# 7 DISCUSSION

In this discussion chapter, we will briefly review the results obtained, aiming towards providing another layer of interpretation that can be absorbed. Furthermore, we will discuss the social impacts that this research can generate.

## 7.1 Arguing the Results

When facing the problem we realized it is difficult to solve it due to the great variety of our classes. The concept of archetype can be very subjective, and in the case of the archetypes used in this work, the concepts can even be mixed. All individuals have at least a little bit of each archetype, which we can see in the PACE report, making it necessary to define metrics and thresholds to define whether a given individual belongs to a certain archetype or not.

PACE appears to be a black box, considering it is a private business tool. In other words, we cannot know exactly how the methodology works to move from adjectives to the archetype. This makes experiments difficult, considering that knowing which marked adjectives influence the decision of the archetype would allow us to do the same using the adjectives present in the text.

The general problem of classifying behavioral profiles using text has proven to be difficult all through out the literature. Even works that use methodologies that exist for decades, and more robust datasets had difficulties creating a good classification model.

This work took several different approaches to extract as much information as possible that could contribute to identifying the archetype. Using multi-class and binary approaches showed that in addition to presenting better results, it is also easier to explain. Within binary approaches was possible to perform regression and classification, using different algorithms, features and textual representations. State-of-the-art techniques were applied in both textual representation and classification, which is the case with transformers, such as BERT and GPT.

Despite this, the best result was obtained when combining the SVM algorithm with tokenization. This combination achieved better performance than using word embedding and transformers for textual representation and classification. Even surpassing the use of GPT fine tuning provided by openAI. Obtaining accuracy above or equal to 63% in all archetypes, with emphasis on the planner profile with an accuracy of 65%.

## 7.2 Social Impacts

The idea of using texts to assess profiles allows for efficiency advantages compared to existing forms of behavior profile assessment, which would require less time than interviews or answering questionnaires. But some doubts can be raised, such as the privacy of the use of curriculum texts, which contain relevant information about an individual, and about automatic technologies replacing human work, which can generate unemployment in the HR area.

When it comes to people's privacy, it is important to point out that our idea is to use texts taken directly from resumes, which is already a document widely used by people to apply for a job. The main objective is also to use already available information, such as job vacancy sites and work focused social networks such as LinkedIn. In addition, the collected data must remain confidential, as it happens when answering a questionnaire or in a job interview, the information is only used to determine the archetype of the individual.

With more and more advances in the areas of technology, especially related to machine learning and artificial intelligence, there is a growing concern that AI-based tools will dominate tasks and replace the people who previously fulfilled that function. The idea of models, such as those proposed in this work, creates the worry of it replacing HR teams, because the models show the potential to be less evasive and more efficient.

It is important to point out that the idea of the model is not to replace, but to be a new tool to add to the existing forms and to be one more way for the HR teams to reach theirs goals. It is clear that older techniques and tools were already losing their space and will continue to lose more and more space, but when it comes to dealing with individuals, we believe that other people are still a necessary part of the process. The main change that this research can make is the need for a new adaptation and learning for those who work in the HR area, to familiarize themselves with the new tools and how to take better advantage of those tools to also facilitate and improve their work.

# 8 CONCLUSION

This work presents a study on the relationship between CV texts in Portuguese and archetypes in the Brazilian organizational environment. Experiments with different approaches were carried out in an attempt to identify these archetypes only with the information obtained from these texts. The task proved to be difficult, as it involves information that is often subjective, as is the case with archetypes, combined with the large amount of information that is textual databases. Despite this, the results obtained show potential in in identification of these archetypes. These models can be very useful, and add a lot to companies' HR teams seeking to improve their organizational environment and team management. In addition to being able to help people understand themselves better, and find their best place to work.

This work explore different existing approaches aiming towards classifying profiles using textual productions, presenting the main works available in the literature. Different techniques and algorithms were applied, and it was evaluated how the metrics behave. Results obtained with a topic modeling experiment provide evidence that there are differences between texts produced by people of different archetypes. However, it is not enough to prove hypothesis 1. This requires more experimentation.

Binary classification experiments prove to be better than multi-class classification and regression. The best experiment in binary classification was combining Tokenization with the SVM algorithm achieved an accuracy close to 65%. These results brought us closer to validating hypothesis 2, however they are still too low to say that there is efficiency in classifying archetypes automatically using text. Experiments with archetypes require complex approaches, and complex features, and shows to be a difficult problem, regardless of the approach chosen. Even if the results do not allow us to confirm the hypothesis, they are still relevant results, considering the difficulties of the problem, and the use of a small, simpler and easily accessible dataset. Considering that people's CVs are available everywhere, and even for free on sites like LinkedIn for example.

Finally, in the explainability experiment, it was noticed that there is noise in the models' decision-making, words that should be used to help define a text of a certain archetype are used for the opposite. Therefore, it was not possible to validate hypothesis 3, the models do not show consistency, and it is necessary to explore the reason for these counter intuitive decisions. Once again, this is most likely due to the difficulty of the problem, which allows each individual to have some of each archetype, in addition to the

simplicity of the texts in the dataset.

For future work there is the possibility of using new databases of CV texts. Furthermore, there is the possibility to use text transcribed from video Curriculums. Another possibility is the combination of textual features with other non-textual features, such as video and audio that can also be extracted from video curriculum.

During the development of this work, two articles with preliminary results were developed, the first (MARIANO et al., 2023) was published at the ICEIS 2023 conference, and the second an extended version awaiting review of the Springer Book of ICEIS 2023.

# REFERENCES

AHONEN-MYKA, H. et al. Applying data mining techniques for descriptive phrase extractionin digital document collections. In: Proceedings of IEEE International Forum on Research and Technology Advances in Digital Libraries -ADL'98-. Santa Barbara, CA, USA: IEEE, 1998. p. 2–11.

ALPAYDIN, E. Introduction to Machine Learning, fourth edition. [S.l.]: MIT Press, 2020. (Adaptive Computation and Machine Learning series).

ANGRAVE, D. et al. HR and analytics: Why HR is set to fail the big data challenge. Human Resource Management Journal, v. 26, n. 1, p. 1–11, jan. 2016.

ARGAMON, S. et al. Lexical predictors of personality type. In: Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America. St. Louis, Missouri, USA: Interface Foundation of North America, 2005.

BASSI, L. Raging debates in HR analytics. Human Resource Management International Digest, v. 20, n. 2, p. 14–18, mar. 2012.

BOJANOWSKI, P. et al. Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, v. 5, p. 135–146, 06 2017.

BROWN, P. F. et al. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, MIT Press, Cambridge, MA, v. 19, n. 2, p. 263–311, 1993.

BROWN, T. B. et al. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. [S.l.]: Curran Associates Inc., 2020. (NIPS '20).

BäCKSTRöM, M.; BJöRKLUND, F.; LARSSON, M. R. Criterion Validity is Maintained When Items Are Evaluatively Neutralized: Evidence from A Full–Scale Five–Factor Model Inventory. European Journal of Personality, Vol. 28, n. 6, p. 620–633, nov 2014.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. an Francisco California USA: ACM, 2016. (KDD'16), p. 785–794.

CHURCHILL, R.; SINGH, L. The evolution of topic modeling. ACM Computing Surveys, ACM New York, NY, v. 54, n. 10s, p. 1–35, 2022.

CORTES, C.; VAPNIK, V. Support-vector networks. Kluwer Academic Publishers, USA, v. 20, n. 3, p. 273–297, sep 1995. ISSN 0885-6125.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv, 2018.

DODGE, J. et al. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. arXiv, 2020.

EIGENSCHINK, M. et al. A critical examination of the main premises of traditional chinese medicine. Wiener klinische Wochenschrift, v. 132, 03 2020.

EYSENCK, H. J.; EYSENCK, S. The eysenck personality inventory. British Journal of Educational Studies, v. 14, n. 1, 1965.

FARNADI, G. et al. Recognising personality traits using facebook status updates. Proceedings of the International AAAI Conference on Web and Social Media, v. 7, n. 2, p. 14–18, Aug. 2021.

GAL, U.; JENSEN, T. B.; STEIN, M.-K. Breaking the vicious cycle of algorithmic management: A virtue ethics approach to people analytics. Information and Organization, Elsevier, v. 30, n. 2, p. 100301, 2020.

GAUR, B.; RIAZ, S. A two-tier solution to converge people analytics into hr practices. In: IEEE. Proceedings of the 2019 4th International Conference on Information Systems and Computer Networks (ISCON). GLA University, Mathura, Uttar Pradesh, India.: ISCON, 2019. p. 167–173.

GILL, A.; NOWSON, S.; OBERLANDER, J. What are they blogging about? personality, topic and motivation in blogs. Proceedings of the International AAAI Conference on Web and Social Media, v. 3, n. 1, p. 18–25, 01 2009.

GILL, A.; OBERLANDER, J. Taking care of the linguistic features of extraversion. In: GRAY, W.; SCHUNN, C. (Ed.). Proceedings of the 24th Annual Conference of the Cognitive Science Society. Fairfax USA: Lawrence Erlbaum Associates, 2002. p. 363–368.

GOLBECK, J. et al. Predicting personality from twitter. In: Proceedings of 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing. MIT, Boston, USA: IEEE, 2011. p. 149–156.

GROOTENDORST, M. R. Bertopic: Neural topic modeling with a class-based tf-idf procedure. ArXiv, 2022.

Guoyin Jiang; Bin Hu; Youtian Wang. Agent-based simulation approach to understanding the interaction between employee behavior and dynamic tasks. Simulation, v. 87, n. 5, p. 407–422, maio 2011.

HEUVEL, S. van den; BONDAROUK, T. The rise (and fall?) of HR analytics: A study into the future application, value, structure, and system support. Journal of Organizational Effectiveness: People and Performance, v. 4, n. 2, p. 157–178, jan. 2017.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. Neural computation, v. 9, p. 1735–80, 12 1997.

JAFARI, R.; FAR, B. H. Behavioral mapping, using nlp to predict individual behavior : Focusing on towards/away behavior. In: Proceedings of 2022 International Conference on Advanced Enterprise Information System (AEIS). London,UK: IEEE, 2022. p. 120–126.

JATNIKA, D.; BIJAKSANA, M.; ARDIYANTI, A. Word2vec model analysis for semantic similarities in english words. Procedia Computer Science, v. 157, p. 160–167, 01 2019.

JOHNSON, S. J.; MURTY, M. R. An aspect-aware enhanced psycholinguistic knowledge graph-based personality detection using deep learning. SN Comput. Sci., Springer-Verlag, Berlin, Heidelberg, v. 4, n. 3, mar 2023.

JUNG, C. G.; HULL, R. F. C. Psychological types. London: Routledge, 1971. (Bollingen series, 6). ISBN 9780415045599.

KARANATSIOU, D. et al. My tweets bring all the traits to the yard: Predicting personality and relational traits in online social networks. ACM Trans. Web, Association for Computing Machinery, New York, NY, USA, v. 16, n. 2, may 2022. ISSN 1559-1131.

KHURANA, D. et al. Natural language processing: state of the art, current trends and challenges. Multimedia Tools and Applications, Springer Science and Business Media LLC, v. 82, n. 3, p. 3713–3744, jul. 2022.

KIM, Y. Convolutional neural networks for sentence classification. In: MOSCHITTI, A.; PANG, B.; DAELEMANS, W. (Ed.). Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014. p. 1746–1751.

LUYCKX, K.; DAELEMANS, W. Personae: a corpus for author and personality prediction from text. In: Proceedings of International Conference on Language Resources and Evaluation. Marrakech, Morocco: ELRA, 2008.

MAIRESSE, F. et al. Using linguistic cues for the automatic recognition of personality in conversation and text. J. Artif. Int. Res., AI Access Foundation, El Segundo, CA, USA, v. 30, n. 1, p. 457–500, nov 2007. ISSN 1076-9757.

MAJUMDER, N. et al. Deep learning-based document modeling for personality detection from text. IEEE Intelligent Systems, v. 32, n. 2, p. 74–79, 2017.

MANNING, C.; SCHUTZE, H. Foundations of Statistical Natural Language Processing. [S.l.]: MIT Press, 1999. (Foundations of Statistical Natural Language Processing).

MARIANO, R. V. R. et al. Natural language processing approach for classification of archetypes using text on business environments. In: Proceedings of International Conference on Enterprise Information Systems. Prague, Czech Republic: SciTePress, 2023. p. 501–508.

MARSTON, W. Emotions of Normal People. [S.l.]: K. Paul, Trench, Trubner & Company Limited, 1928. (International library of psychology, philosophy, and scientific method).

MCDOUGALL, W. Of The Words Character and Personality. Journal of Personality, Vol. 1, n. 1, p. 3–16, set. 1932.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. In: Proceedings of International Conference on Learning Representations. [S.l.: s.n.], 2013.

MORENO, J. D. et al. Can personality traits be measured analyzing written language? a meta-analytic study on computational methods. Personality and Individual Differences, v. 177, p. 110818, 2021.

NADEAU, D.; SEKINE, S. A survey of named entity recognition and classification. Lingvisticae Investigationes, v. 30, 08 2007.

NOWSON, S.; OBERLANDER, J. Identifying more bloggers: Towards large scale personality classification of personal weblogs. In: Proceedings of International Conference on Web and Social Media. Boulder, USA: ICWSM, 2007.

OBERLANDER, J.; NOWSON, S. Whose thumb is it anyway? classifying author personality from weblog text. In: Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions. Sydney, Australia: Association for Computational Linguistics, 2006. p. 627–634.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. Found. Trends Inf. Retr., Now Publishers Inc., Hanover, MA, USA, v. 2, n. 1–2, p. 1–135, jan 2008.

PARK, G. J. et al. Automatic personality assessment through social media language. Journal of personality and social psychology, v. 108 6, p. 934–52, 2015.

PENNEBAKER, J.; KING, L. Linguistic styles: Language use as an individual difference. Journal of personality and social psychology, v. 77, p. 1296–312, 1999.

PENNEBAKER, J. W.; MEHL, M. R.; NIEDERHOFFER, K. G. Psychological aspects of natural language use: Our words, our selves. Annual Review of Psychology, v. 54, n. 1, p. 547–577, 2003.

PENNINGTON, J.; SOCHER, R.; MANNING, C. GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543.

PETERS, M. E. et al. Deep contextualized word representations. In: WALKER, M.; JI, H.; STENT, A. (Ed.). Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). [S.l.]: Association for Computational Linguistics, 2018. p. 2227–2237.

RADFORD, A.; NARASIMHAN, K. Improving language understanding by generative pre-training. Computer Science, Linguistics, 2018.

RAGUVIR, S.; BABU, S. Enhance employee productivity using talent analytics and visualization. In: Proceedings of 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI). Sakheer, Bahrain: IEEE, 2020. p. 1–5.

RAJPURKAR, P. et al. SQuAD: 100,000+ questions for machine comprehension of text. In: SU, J.; DUH, K.; CARRERAS, X. (Ed.). Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistics, 2016. p. 2383–2392.

RAMLAWATI, R. et al. External alternatives, job stress on job satisfaction and employee turnover intention. Management Science Letters, v. 11, n. 2, p. 511–518, 2021.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 1135–1144.

SALTON, C. B. G. Term-weighting approaches in automatic text retrieval. Information Processing and Management, v. 24, n. 5, p. 513–523, 1988. ISSN 0306-4573.

SANTOS, V.; PARABONI, I.; SILVA, B. Big five personality recognition from multiple text genres. In: Proceedings of International Conference on Text, Speech and Dialogue. Prague, Czech Republic: Springer, 2017. p. 29–37.

SANTOS, W. R. dos; PARABONI, I. Personality facets recognition from text. In: Proceedings of Experimental IR Meets Multilinguality, Multimodality, and Interaction. Cham: Springer International Publishing, 2019. p. 185–190.

SHAPIRO, E. S.; BROWDER, D. M. Behavioral Assessment. In: MATSON, J. L. (Ed.). Handbook of Behavior Modification with the Mentally Retarded. Boston, MA: Springer US, 1990, (Applied Clinical Psychology). p. 93–122.

SMITH, B. L. et al. Effects of speech rate on personality perception. Language and Speech, v. 18, n. 2, p. 145–152, 1975.

TURSUNBAYEVA, A.; LAURO, S. D.; PAGLIARI, C. People analytics—a scoping review of conceptual boundaries and value propositions. International Journal of Information Management, Elsevier, v. 43, p. 224–247, 2018.

VASWANI, A. et al. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems. Long Beach, USA: Curran Associates, Inc., 2017. v. 30.

VIEIRA, A. G. et al. A probabilistic mapping approach to assess the employee behavior profile. In: Proceedings of 2023 IEEE 25th Conference on Business Informatics (CBI). Prague, Czech Republic: IEEE, 2023. p. 1–8.

VU, X.-S. et al. Lexical-semantic resources: yet powerful resources for automatic personality classification. arXiv, 2017.

WABER, B. People analytics: how social sensing technology will transform business and what it tells us about the future of work. [S.l.]: FT Press, 2013.

ZHAO, J. et al. User personality prediction based on topic preference and sentiment analysis using lstm model. Pattern Recognition Letters, v. 138, p. 397–402, 2020.

ZHU, Y. et al. A lexical psycholinguistic knowledge-guided graph neural network for interpretable personality detection. Knowledge-Based Systems, v. 249, p. 108952, 2022.