PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
Programa de Pós-graduação em Engenharia Elétrica

Cibele Simões de Oliveira Santos

CVO: CURRICULUM VITAE OPTIMIZATION BY RECOMMENDING KEYWORDS TO UNDERGRADUATE STUDENTS

Belo Horizonte
2021

Cibele Simões de Oliveira Santos

CVO: CURRICULUM VITAE OPTIMIZATION BY RECOMMENDING KEYWORDS TO UNDERGRADUATE STUDENTS

Dissertação apresentada ao Programa de Pós-graduação em Engenharia Elétrica da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para obtenção do título de Mestre em Engenharia Elétrica.

Orientador: Prof. Dr. Luís Fabrício Wanderley Góes

Coorientador: Prof. Carlos Augusto Paiva da Silva Martins

Belo Horizonte
2021

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

Santos, Cibele Simões de Oliveira
S237c CVO: curriculum vitae optimization by recommending keywords to undergraduate students / Cibele Simões de Oliveira Santos. Belo Horizonte, 2021.
55 f.: il.

Orientador: Luís Fabrício Wanderley Góes
Coorientador: Carlos Augusto Paiva da Silva Martins
Dissertação (Mestrado) – Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Engenharia Elétrica

1. Curriculum vitae. 2. Estudantes universitários. 3. Pessoal – Recrutamento. 4. Sistemas de recuperação da informação. 5. Processo decisorio - Processamento de dados. 6. Processamento de linguagem natural (Computação). 7. Algorítmos. 8. Aprendizado do computador. 9. Support vector machines. I. Góes, Luís Fabrício Wanderley. II. Martins, Carlos Augusto Paiva da Silva. III. Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Engenharia Elétrica. IV. Título.

CDU: 681.3.091

Ficha catalográfica elaborada por Fabiana Marques de Souza e Silva - CRB 6/2086

Cibele Simões de Oliveira Santos

CVO: CURRICULUM VITAE OPTIMIZATION BY RECOMMENDING KEYWORDS TO UNDERGRADUATE STUDENTS

Dissertação apresentada ao Programa de Pós-graduação em Engenharia Elétrica da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para obtenção do Título de Mestre em Engenharia Elétrica em Programa de Pós-graduação em Engenharia Elétrica.

Prof. Dr. Luís Fabrício Wanderley Góes (Orientador) - PUC Minas

---

Prof. Carlos Augusto Paiva da Silva Martins (Coorientador) - PUC Minas

---

Prof. Dr. Felipe Domingos da Cunha - PUC Minas

Profa. Dra. Milene Barbosa Carvalho - UFSJ

Belo Horizonte, 30 de Julho de 2021

To my family and friends, for all the love and support during my master's degree journey.

ACKNOWLEDGEMENTS

I would first like to thank to PUC Minas, for my training as a systems analyst and training as a master.

Second, I would like to thank to my advisors, Prof. Dr. Luís Fabrício Wanderley Gomes, Prof. Dr. Felipe Domingos da Cunha and Prof. Dr. Carlos Augusto Paiva da Silva Martins, for having chosen me, guided me, for the knowledge they conveyed to me, availability, patience and friendship.

I would like to acknowledge to the graduate professors for their knowledge, availability, patience, friendship and support during difficult times. I would particularly like to single out members of the PPGEE secretariat, especially to Eliza Silva Caetano, for her assistance in all matters related to the graduate program. To my friends and graduate colleagues, for their friendship and help. It was an honor to live with electrical engineers and share knowledge.

I would also like to thank my parents, my mother Elizete Simões Santos and my father Celso de Oliveira Santos, for being an example, for their teachings, affection and love. To my brothers Junio, Filipe and Jessica always worrying about my mental health, for all the moments of comfort and joy. To my fiance, Sérgio Vitarelli Silva for his patience, dedication and support at all times. To my friends and family for their friendship and help in difficult times.

"We can only see a short distance ahead, but we can see plenty there that needs to be done." - TURING

ABSTRACT

Candidate selection platforms have been widely used in companies that seek agility in the process of hiring. Candidates who do not meet the requirements of a job vacancy are disqualified in the first step, called screening. This stage has been automated due to the large volume of curriculum vitae (CV) of candidates per vacancy, particularly for internship vacancies. As a consequence, candidates receive little to none feedback and do not know how to improve/optimize their CVs for new applications. The goal of this dissertation is to develop the curriculum vitae optimization (CVO) process for internship vacancies by implementing a recommendation system that given an undergraduate student CV, it suggests the addition of relevant keywords, taking into account the student's undergraduate course. This system is implemented based on the clustering of CVs keywords, from an internship recruitment private company database, into profile groups which are linked to internship vacancies. The experimental results showed that recommendations improved students CVs similarity (competitiveness within a specific field) from 18.83%, with 3 keywords recommendation, up to 50.67%, with 10 words.

Keywords: Clustering; Curriculum vitae optimization; Keywords recommendation.

RESUMO

As plataformas de seleção de candidatos têm sido amplamente utilizadas em empresas que buscam agilidade no processo de contratação. Os candidatos que não atendem aos requisitos de uma vaga de emprego são desclassificados na primeira etapa, chamada de triagem. Essa etapa vem sendo automatizada devido ao grande volume de curriculum vitae (CV) de candidatos por vaga, principalmente para vagas de estágio. Como consequência, os candidatos recebem pouco ou nenhum feedback e não sabem como melhorar/otimizar seus currículos para novas inscrições em vagas. O objetivo desta dissertação é desenvolver o processo de otimização do curriculum vitae (CVO) para vagas de estágio por meio da implantação de um sistema de recomendação que, dado o currículo de um aluno de graduação, sugere a adição de palavras-chave relevantes, levando em consideração o curso de graduação do aluno. Este sistema é implementado com base no agrupamento de palavras-chave de CVs, a partir de uma base de dados de um empresa privada de recrutamento de estágios, em grupos de perfis que estão associados às vagas de estágio. Os resultados experimentais mostraram que as recomendações melhoraram a similaridade dos currículos dos alunos (competitividade dentro de um campo específico) de 18,83%, com recomendação de 3 palavras-chave, para 50,67%, com 10 palavras.

Palavras-chave: Clustering; Otimização de curriculum vitae; Recomendação de Palavras-chave.

LIST OF FIGURES

LIST OF TABLES

SUMMARY

1 INTRODUCTION

1.1 Context

The undergraduate student career starts with an internship, where they can practice the skills learnt at the university. In this context, there is an increasing competition for internship vacancies, in which students are first evaluated by their curriculum vitae (CV) (PIERRE; JEANNE, 2020). However, crafting a CV is not an easy task particularly in online recruitment, where each CV can be seen as a webpage and requires the careful use of relevant keywords to be better ranked in recruitment search engines such as Linkedin. This process is similiar to SEO (search engine optimization) techniques for improving websites ranking on search engines, we thus call the process of improving CVs in this context as CVO (curriculum vitae optimization) (RAMANATH et al., 2018). Most online services for internship applications are focused on making it easier and efficient for companies recruiters to identify potential candidates, instead of helping undergraduate students to improve their CVs, increasing their odds to be invited for an internship interview.

This problem has been identified in previous work such as (DING et al., 2017), the majority of recommendation systems focus on job search services and simple recommendations, ignoring the specific needs of undergraduates seeking for a first job opportunity. Those students usually do not have any professional experience and the required skills for a full job position. One way used by authors was to purpose a system that uses the student grade in university combined with machine learning algorithms to find the best match between student and internships. On the other hand, to find similarity between profiles, (RODRIGUEZ; CHAVEZ, 2019) focuses his research on calculating the similarity between two profiles where values of common characteristics are extracted and clustered for a specific job position. Both approaches are aimed at classifying skills according to the job position, rather than providing feedback for candidates on how to improve their CVs.

According to (NGO, 2019) the recruitment is defined by a process in which companies create vacancies to job positions in order to select appropriate candidates. It includes the establishment of job positions in many websites to select candidates by their curriculum vitae. The CVs summarizes a candidate's career and qualifications and is a mechanism to provide relevant personal and business characteristics to a potential employer (DIYA, 2003). The CV usually includes the candidate's career goal, personal interests, professional affiliations, educational background, job history, and a description of work experience. In this way, companies have a chance to recruit suitable and potential employees through universities (NGO, 2019). The students are those who are constantly looking for job op-

portunities in order to consolidate the knowledge acquired during the university in a period called internship. Interns training benefits companies in the long term because, after graduation, they have a greater chance of returning to internship companies to work full-time.

1.2 Motivation

There are many students who do not have someone to help them improve their curriculum in Computing field for example. Otherwise, as quoted by (WANG et al., 2020), there are some problems that the graduate students confronts:

1. Universities lack of ways to understand the specific situation of the current employment market.

2. Universities cannot obtain feedback from the job market and sometimes courses and skills learned by students cannot meet the needs of enterprises.

3. There are many false information and messy information in online recruitment, the graduates have not a clear view about their career pursuit because of the relatively narrow employment channels.

In this way, the main motivation of this dissertation is the curriculum vitae optimization (CVO) to help the graduation students suggesting the words most common in CVs that fits in jobs positions in their areas, in order to them achieve the internship vacancy.

1.3 Objective

The goal of this dissertation is to develop the curriculum vitae optimization (CVO) process for internship vacancies by implementing a recommendation system that given an undergraduate student CV, it suggests the addition of relevant keywords, taking into account the student's undergraduate course.

This system is implemented based on the clustering of CVs keywords, from an internship recruitment private company database, into profile groups which are linked to internship vacancies. It enables undergraduate students to improve their CVs using skill recommendations and feedback to achieve effectiveness in an internship job. The profile

groups were able to cluster students into 9 groups, covering the most popular undergraduate courses on this dataset and the recommendations improved students CVs similarity (competitiveness within a specific field) from 18.83% with 3 keywords recommendation.

1.4 Dissertation structure

The remaining of this dissertation is organized as follows. In Section 2, the Background is presented. In section 3, there is the Related Work. In section 4, there is the Methodology. Next, Section 5 presents and analyzes the Experimental Results. This dissertation is concluded in Section 6.

2 BACKGROUND

2.1 Recruitment for Internships

As quoted by (NGO, 2019), recruitment is defined as a process in which companies create job openings for positions in order to select the appropriate candidates. This includes establishing jobs on different websites for qualified and suitable candidates to apply for positions. This process is usually carried out through resumes.

According to (DIYA, 2003), the resume summarizes a candidate's career and qualifications. A resume is a mechanism for conveying personal and business characteristics that the candidate believes been relevant to a potential employer. The resume usually includes the candidate's career objective, personal interests, professional affiliations, educational background, employment history, and a description of work experience.

Companies have the chance to recruit suitable and potential employees through universities (NGO, 2019). Students are those who are constantly looking for work opportunities in order to consolidate the knowledge acquired during university in a period called internship. Intern training benefits companies in the long term as, upon graduation, they have a greater chance of returning to internship companies to work full-time or actually be hired.

2.2 Data Science

According to (WALLER; FAWCETT, 2013), Data Science is the application of quantitative and qualitative methods to solve relevant problems and predict results, putting data in perspective. This area can also be understood as a set of techniques to extract and analyze data from a statistical point of view, aligned with the use of machine learning algorithms. These algorithms can be used in a supervised, unsupervised or reinforcement way, they are capable of interpreting large amounts of data and, as an output, group and classify similar information from a given set of characteristics.

In the analysis proposed in (KAUR, 2019), the steps of Data Science for decision making are presented. First, it is necessary to establish what the research objective is and how those involved benefit from this research. The next step after completing the research objective is to collect the data. There can be many errors in the data during the collection process. Once the collection is finished, the data preparation is carried out, where the data undergoes cleaning, integration and transformation before being applied

in the data model. Data exploration provides a deeper understanding of the collected information.

During data exploration, data scientists try to understand how variables communicate with each other and how data is distributed, as well as better insight into the data. When data is combined with domain knowledge, you can use it to create data models. Building a model is an iterative process that doesn't just select one variable. This one is run several times for productive diagnostics. The last step is to present the results in a generally automated model so that the information acquired can be used by those involved.

2.3 Natural Language Processing

According to (CAMBRIA; WHITE, 2014), Natural Language Processing (*Natural Language Processing - NLP*) is a range of theoretically motivated computational techniques for automatic analysis and representation of human language. Since its inception in the 1950s, NLP research has focused on tasks such as machine translation, information retrieval, text summarization, answering questions, information extraction, topic modeling, opinion mining and others.

For a machine to understand that the words on the resume are actually describing a candidate's abilities, it needs to be guided in interpreting the data. Therefore, NLP techniques are used for a better understanding and interpretation of key terms in the CV. In order to use NLP to experiment with the texts of the CV in this research, the steps of pre-processing texts in NLP are carried out. Are they:

1. *Normalization*: the text is transformed into tokens, that is, vectored for treatments such as cleaning, correction of upper and lower case letters and removal of special characters;

2. *Removal of Stopwords*: in this step, pronouns, articles and words that do not necessarily dictate user characteristics are removed;

3. *Spelling Correction*: in this step, names of university courses and accentuation are corrected;

4. *Stemming*: finally, the semantic parts of the words are obtained, in order to facilitate the lexicographic classification of terms.

2.4 Clustering Tecniques

Text clustering is an unsupervised machine learning technique (LIU et al., 2020). For better clustering of text, the terms within a cluster are considered the most similar terms, while the texts between the clusters are the most different. This means that the performance of text clustering depends on both the similarity measurement and the grouping. The clustering, results in groups of objects through common characteristics where it is possible to find the arrangement of the representation of the words in the three-dimensional plane and also their similarity.

Among the algorithms used for clustering, one of the most common is K-Means where initially, the algorithm selects a randomly chosen center for each cluster and assigns each data in the training set to one of the cluster whose center is closest. The algorithm recalculates the center of the clusters and continues until there is no significant change in the center value. The K-means algorithm works on the assumption that all attributes are independent and normally distributed (SIVARAM; RAMAR, 2010).

Obtaining the optimal number of clusters can be done in several ways, with two of the most common are: the *Ebow-curve* and the *Silhoette Method*.

According to the authors in (YUAN; YANG, 2019), the basic idea of the *Ebow-curve* is to use a square of the distance between the sample points in each cluster and the cluster centroid to provide a series of K values. The Sum of Squares Errors (*Error Sum of Squares - SSE*) is used as a performance indicator. The K value is repeated and the SSE is calculated. Smaller values indicate that each cluster is more convergent. When the number of *clusters* is set to approximate the number of real clusters, SSE shows a rapid decline. When the number of clusters exceeds the number of actual clusters, the SSE will continue to decrease, but it will quickly become slower.

According to (YUAN; YANG, 2019), the Silhouette method * combines the two factors of cohesion and resolution. Cohesion is the similarity between the object and the cluster. When compared to other clusters, it is called separation. This comparison is obtained by the Silhouette value, which is in the range of -1 to 1. The Silhouette value of close to 1, indicates that there is a close relationship between the object and the cluster.

* The Silhouette method was first proposed by J. Rousseeuw (YUAN; YANG, 2019). First, the average distance $a(i)$ from sample i to other samples in the same cluster is calculated. The smaller $a(i)$ is, the more the i sample must be clustered in the cluster. $a(i)$ is referred to as the intra-cluster dissimilarity of the i sample. The average $a(i)$ of all samples in cluster c is called cluster dissimilarity c. Second, the average distance $b(i)$ of all samples in sample i to the other cluster, $c(i)$ is calculated, which is called the dissimilarity between sample i and cluster $c(i)$. Defined as the inter-cluster dissimilarity of sample i: $b(i) = minbi1, bi2, ..., bik$ the larger $b(i)$, the less sample i belongs to others clusters. Finally, the boundary coefficients of the sample i are defined according to the intra-cluster dissimilarity $a(i)$ of the sample i and for the inter-cluster dissimilarity $b(i)$ (YUAN; YANG, 2019).

If a cluster of data in a model is generated with a relatively high Silhouette value, the model is suitable and acceptable.

2.5 Classification Techniques

According to (DURAIRAJ; VIJITHA, 2014), data classification is the categorization of data for its most effective and efficient use. In a basic approach to storing computer data, data can be classified according to its critical value or how often it needs to be accessed, with the most critical or frequently used data stored on the fastest media, while other data can be stored more slowly (and less expensive). This type of classification tends to optimize the use of data storage for various purposes, not just relative importance or frequency of use but technical, administrative, legal and economic.

Taking into account classification of profiles, this technique is used to categorize the user according to the keywords of their resume. For this classification, words are calculated by frequency and then labeled according to the similarity cluster. Thus, this dataset with labels can be used for training, where, according to the calculated characteristics, a new dataset can be inserted into this model and then labeled.

Several techniques can be used for such classification as Decision Tree, SVM, Random Forests, among others. Following we present a description of the main ones:

2.5.1 Decision Tree

This algorithm is a tree structure where the internal nodes denote a test on an attribute, the branches represent the test results and the leaf nodes represent the labels of the (SIVARAM; RAMAR, 2010) classes. Regarding health, the algorithm creates a node and then applies the attribute selection method to determine the best splitting criteria and the created node is named by that attribute. The training tuple subset is formed using the *splitting* attribute. The algorithm is called recursively for each subset, until the subset contains tuples of the same class. When the subset contains tuples of the same class, a leaf is appended with a majority class label in the root training set.

2.5.2 KNN

The *K-Nearest Neighbors (KNN)* kNN algorithm is a non-parametric classification method that indicates how similar a vector is to another vector. First, the KNN of a given test sample from the entire set of training samples available is determined. Neighbors are determined based on Euclidean distance (may also be Manhattan, Minkowski, or Weighted). The distribution of classes of these neighbors is observed. The final decision is made based on the class that provides the most appearances among the results that had the smallest distances. (NARAYANAN; DJANEYE-BOUNDJOU; KEBEDE, 2016).

2.5.3 Naive Bayes

This algorithm assumes that the value of a given resource is not related to the presence or absence of any other, given the class value. This one considers that each of these characteristics independently contributes to the probability that the record actually belongs to the class, regardless of the presence or absence of the other characteristics. An advantage of Naive Bayes is that it requires only a small amount of training data to estimate the parameters needed for classification. In this case, the classification of the Naive Bayes algorithm may be measured using a confusion matrix (also known as an error matrix) that allows measuring the performance of a classification algorithm. As the independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire confusion matrix (DURAIRAJ; VIJITHA, 2014).

2.5.4 Random Forest

The *Random Forest*, or Random Forests, algorithm, first introduced by Breiman in the early 2000s, represents a widely applied approach to pattern recognition (BICEGO, 2019). This technique represents a set of decision trees, a widely known classification tool that performs a hierarchical division of the resource space, each division being based on the boundary of a single resource. The algorithm uses the "divide and conquer" principle: in its classic version, different sets of random samples are extracted from the training set, each used to build a decision tree. The final model is then obtained by aggregating the different trees. This aggregation shows interesting theoretical properties, the most famous being that shown by Breiman, who defined an upper bound for the generalization error of Random Forests in terms of correlation and strength of individual trees. The key aspect

of *Random Forest* is represented by the randomization injected in learning the model, which allows to have strongly diverse trees. The preference for this algorithm is due to the fact that a large number of trees operating together will perform better than other individual constituent models.

2.5.5 SVM

The *Support Vector Machines (SVM)* algorithm is a discriminative type of classification technique. The boundaries of each class are well defined by a separation hyperplane where it basically creates an ideal hyperplane that limits the given training dataset and ranks the test sample based on the plane it is in. Various types of kernels can be used for final decision (NARAYANAN; DJANEYE-BOUNDJOU; KEBEDE, 2016).

2.6 Recommendation Systems

Recommendation Systems were developed through a set of tools and techniques that give suggestions to a user about specific items that can be of their interest (KANWAL et al., 2021). According to (ZHENG et al., 2012) in their research, the concepts of recommendation systems involve: the Target User, the Desired Profile and the Similarity between these entities. Once these entities are identified, the research highlighted three different types of recommendation systems such as *Collaborative Filtering Recommenders (CFRs)*, *Content-Based Recommenders (CBRs)* and *Knowledge-Based Recommenders (KBRs)*, which is the concept of development of a recommendation system.

When we talk about text recommendation that intend to recommend skills, the CFRs are the best approach. There is a field inside Recommendation Systems that studies different techniques such Neural Networks to the specific recommendation to profiles. These techniques are used in suggestion systems of companies such as Amazon and Netflix that utilize item attributes and user reviews for this particular purpose (KANWAL et al., 2021).

In order to calculate the similarity between profiles, there are several ways to find it such as: Linked data semantic distance, Euclidean (WEI; VARSHNEY; WAGMAN, 2015) distance, Jaccard similarity and the most famous Cosine Similarity (KANWAL et al., 2021), which will be addressed in this dissertation.

3 RELATED WORK

3.1 Candidate Skills Selection

During the candidates selection process for vacancies, it is necessary to evaluate and measure the aspects that represent the difference among the candidates. These aspects are found during the process of selecting feature characteristics. According to (HARRIS, 2017), the skills selection of a candidate for a vacancy is determined by two aspects: the technical aspect, known as hard skills, and by the personal aspect, known as soft skills. This second determines the motivational aspects of an individual, generally ignored in automated systems, since such systems correlate job requirements with candidates hard skills. However, looking for these characteristics, recruiters use recommendation platforms to group characteristics in job positions. Although the authors at (HARRIS, 2017) have innovated using the concepts of gamification and crowdsourcing, in their work to rank candidates according to job requirements, selecting deterministic characteristics, such as Sex and Age, reduced the diversity of the database.

The work of (RODRIGUEZ; CHAVEZ, 2019) emphasizes the importance of pre-processing the characteristics before associating the candidate with others. The authors highlighted three steps for the correct selection of profiles. The first was screening, removing insignificant records, in some cases with too many missing values or with little variation to be useful. The second was the classification of attributes, allocating the established importance of data entries. Finally, the third is the selection that identifies the cluster of resources to apply subsequent models. With these steps, the reduction of noise in the data was shown, improving mining performance, such as predictive precision and comprehensibility of the results. This pre-processing allows to focus on the relevant parts of the data and improve the capacity of the algorithm. The paper focused mainly on job positions and not candidates. The Weka software was used on a private dataframe to understand which were the most relevant attributes in the selection of job positions such as job Title and work experience.

In (WEI; VARSHNEY; WAGMAN, 2015), a platform was created for internal transfer of professionals in positions at IBM, statistically measuring employees who could or could not be transferred from their positions based on the keywords of their CVs. This search was made by using the experience taxonomy of the positions with the requirements of the jobs and then these were ranked through the Euclidean distance according to the level of demand in the new job. The taxonomy refers to collections that are ordered by a classification scheme and generally organized hierarchically. The skills were presented to the candidates to find which area they found themselves. Although they used keywords

of skills like “Cisco certification" to correspond the employee with the job positions, these authors did not use any techniques to calculate the correspondence between profiles.

A taxonomy was also used by (DIABY; VIENNET, 2014) that created a recommendation system of job positions using profile data from Facebook and LinkedIn. They used the statistical technique of frequency inversion TF-IDF (Term Frequency – Inverse Document Frequency) combined with the SVM algorithm. The authors said that this method did not behave very well in Facebook data. The difference between this research and the author's search is the application of the method in a private data, but also use the TF-IDF method and a comparison with SVM algorithm.

In the selection of characteristics, the difference of this dissertation among the others is the focus of intern profiles that in majority do not have pre-defined experiences. The database diversity is also considered to avoid bias, so characteristics such sex and age are not established in the private database used on this dissertation. The remarkable steps in (RODRIGUEZ; CHAVEZ, 2019) are relevant in this research, with the difference of the focus being not vacancies but intern profiles. Finally, statistical measures are also used as applied on (WEI; VARSHNEY; WAGMAN, 2015) and (DIABY; VIENNET, 2014).

3.2 Clustering of CVs

The second step after selecting characteristics is to find similarities between profiles that share common characteristics. This similarity can be achieved through the technique known as clustering (RODRIGUEZ; CHAVEZ, 2019).

The work proposed in (GUPTA; GARG, 2014) used clusters of user categories based on resumes attributes, and these clusters were used later in a recommendation system developed in the work. A disadvantage of this type of clustering would be the imbalance of a certain cluster in the face of limitations of intervals that may have no data for example, since there are only manual calculations.

In the case of professionals looking for internship positions, the distance between company and the university can be very large. In order to solve this, the research of (DING et al., 2017) used the student historical academic data to find intern positions according to the grades of students at university. As a result, the RRSGR (Reciprocal Recommender System for Graduates' Recruitment) system that reads the academic data of the university candidate for the position, evaluates the profile and presents positions opened by companies in this same system. The work proposes an algorithm that involves undergraduates, alumni, graduates, and the company. The proposed algorithm involving probabilistic neighborhood selection and clustering of priority k-Medoids are used to help

improve the accuracy and diversity of recommendation results. Then a K-Medoids algorithm (a K-Means partition) is proposed as part of this solution. In (SIVARAM; RAMAR, 2010), clustering algorithms such as K-Means, fuzzy C-Means and classification were used to understand which attributes are most determinant in selection of recently graduated candidates. The results showed that classification algorithms like Decision Tree, had better accuracy than the other algorithms when evaluated separately. However the combination of clustering algorithms with classification algorithms was not evaluated for profiles.

In contrast, the research proposed by (DURAIRAJ; VIJITHA, 2014) combined the use of classification and clustering algorithms to predict academic performance of students and their consequences. According to the authors the classification algorithms categorize the data for a better use, it means, the use of this techniques improves educational performance and assessment of the student's learning process. The clustering algorithm used was K-means and for classification, the Decision Tree and Naive Bayes were used. The results were positive for combine the two types of algorithms, and they were evaluated by the Confusion Matrix and ROC curve to verify which students would possibly have better academic performance in the exams. The authors highlighted the importance of these techniques for the decision-making process.

In clustering of CVs, the contribution of this research compared to other works is the identification of the profiles clusters based on the classification of the characteristics that generate the clustering not only of the approved candidates, but also the similarity of the skill set of the candidates profile. Therefore, a cluster that generates a large distance from the profile to the internship position through the keyword trend, should be shared with the candidate for possible change or adaptation of this tendency CV.

3.3 Recommendation Systems

Using Content-Based recommendation systems, in (ALMALIS et al., 2016), the objective of the research was the recommendation of candidates for job positions by quantifying the suitability of a candidate for a job through intervals. They used structured and unstructured data in which each candidate received a score for each of the required qualifications. Finally, according to the qualification, the candidates were indicated or not to the job position using the distance of Minkowski. In order to find better accuracy in skill recommendation, some systems use the combination of more than one machine learning model creating Hybrid Recommendation Systems (*Hybrid Recommenders (HR)*). In (MAURYA; TELANG, 2018), were identified necessary skills for allocating people in jobs using the combination of the algorithms BayesMatch and NpBayesMatch models as recommendation algorithms and associating with other members.

Similarly to the social networks, the process of recommending job positions is calculated by the common characteristics of members, as social networks tend to recommend similar profiles. In (CHALA; FATHI, 2017), the importance of social networks is highlighted in order to identify methods that measure the skills of a candidate for a job position. Still using social networks, the research of (HONG; ZHENG; WANG, 2013) dynamically increases the user's profile, according to the user's history in his search for vacancies on social networks, also using the combination of two recommendation algorithms. However, recommendation systems tend to suggest profiles that are closer to the qualifications of a job position. In some cases, if the candidate does not have a percentage of requirements, the vacancy is not recommended, and the candidate doesn't get a feedback. In the work proposed by (CHALA; HARRISON; FATHI, 2017), the authors focused in necessary and desirable skills (also referred to preferred skills) that are not identified by recommendation systems, such as extra grade. Their research showed how plain keyword-based vacancy-to-jobseeker matching in may result in improper matching, which means that is it important to recommend skills that make sense with the candidate's CV.

In order to solve the problem of university students' employment selection, (WANG et al., 2020) created a text recommendation system to suggest job positions to candidates based on the words of CV. The approach used web crawler techniques, the TF-IDF algorithm, K-Means, and the algorithm LDA (Latent Dirichlet Allocation). The authors showed the top 30 terms that occur in CVs and the relevance of these terms extracting employment skills using the LDA theme model. The combination of these techniques was positive especially when the systems suggest the top terms to the candidate. In another way, the approach purposed by (ROY; CHOWDHARY; BHATIA, 2020), classify the resume in classes of CVs, comparing different algorithms, and ranking the candidate resume based on the job description and their resume content. Their algorithm recommends resumes based on the similarity index with the given job description. Although it classified the candidate's profile based on their resume, this approach also focuses on the recommendation process, not on the candidate itself.

In recommendation, the works are focused on the indication of vacancies, not worrying about the recommend of the profile of a candidate that didn't pass in a process or was not indicated for a vacancy for example. In this work we intend to characterize the groups in order to evolve them where, in addition to the recommendation of skills to intern positions, we also provide feedback, based on the keywords of the curriculum related to the *hard skills* of the candidates.

The table 1 summarizes the most relevant aspects found in the literature according to each theme and the respective contribution to this research classified by main theme.

Table 1 – Resume of relevant aspects found in the literature.

Characteristic	Main Theme	Referência
Candidate Skills Selection	Difference between hard skills and soft skills. It uses a non-diverse base.	Harris (2017)
	Importance of data pre-processing	Rodriguez; Chavez (2019)
	Words like "Cisco Certified" were considered but ML was not used to calculate profile similarity and matching.	Wagman et al. (2015)
	Use of the TF-IDF technique and the SVM algorithm.	Diaby; Vinnet (2014)
Clustering of CVs	Clustering and classification algorithms used to understand attributes.	Sivaran; Ramar (2010)
	Use of University History.	Ding et al. (2017)
	Combination of clustering and classification algorithms. Positive results were obtained.	Durairaj; Vijitha (2014)
Recommendation Systems	Hybrid Recommendation Systems based on social media content.	Almalis et al. (2016)
	Combination of BayesMatch and NpBayesMatch models as recommendation algorithms and associating with other members.	Maurya; Telang (2018)
	Suggest positions to candidates based on the words of the CV. Web crawler techniques, the TF-IDF algorithm, K-Means and the LDA algorithm.	Wang et al. (2020)
	Content-based recommendation rating (KNN) the candidate resume based on job description and resume content. Cosine Similarity Index.	Roy et al. (2020)

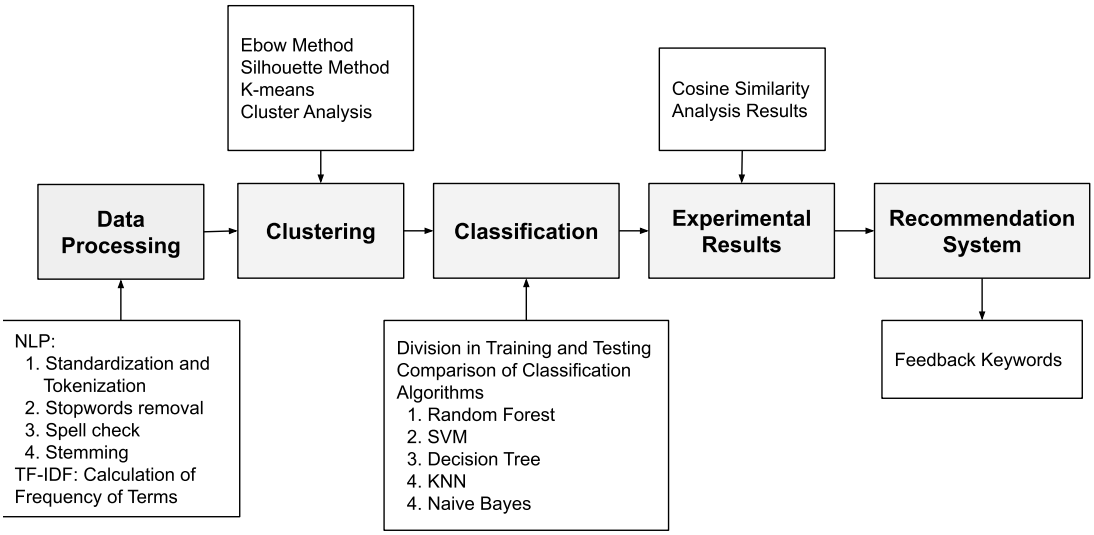
Source: Made by the author.

4 PROPOSED METHOD

4.1 Methodology

This methodology section presents the techniques, database, and tools used for clustering the profile groups by keywords and to build the recommendation system. In the next sections each step will be presented according to the flowchart presented in the Figure 1.

Figure 1 – Methodology steps for approach proposed.



Source: Made by the author.

4.2 Dataframe

The analysis of this research is focused on the keywords of undergraduate students CVs focused on internship vacancies. The private dataset consists of a database with user identifiers, courses, cities and keywords previously 163,885 undergraduate students CVs from Brazilian universities (the database is in Portuguese) available in JSON format by an internship recruitment company. The keywords were previously extracted by the company that provided the dataset.

4.3 Data processing

In the NLP stage, the preprocessing of the data was made. During this preprocessing, the words were unified by the user, creating a single record for each user. After this unification, the normalization step started.

The normalization step, in addition to the removal of columns that contained duplicate information, the vectorization process allowed the cleaning of words, removing special characters and null records. In this stage, words like “turno"(shift) and “habilidades comportamentais" (behavioral skills) were also removed since the focus is on hard skills. Words related to the courses have also been removed as the objective is to seek as many skills as possible. In this step, the was reduced to 7,202 candidates/CVs (rows) and 3 columns: User ID, Course and Standardized Keywords. The second stage of data processing was the removal of stopwords, which are words that are considered irrelevant in a search. For example: “do”, “da”, “os”; generally used in names of courses and technologies such as “computer science”, in this case “da”, is a *stopword*. In this step the Python NLKT library was used, and then all the stopwords from the Portuguese dictionary were imported. The functions below were used to remove *stopwords* for each record of each user: $remove\_stop\_words()$, $remove\_pt\_br\_char\_by\_text()$ and $replace\_ptbr\_char\_by\_word()$.

In the third step, stemming was performed, that is, given the set of words from the *corpus*, the lexicographic parts of the terms are obtained, in short, this generalizes the words transforming the set of each user, to a set of words in common with other users. For example, the words “engenheiro”, “engenharia”(engineer), “engenheira” have the same radical “engenh”, so the algorithm transforms these words into this single radical.

4.3.1 TF-IDF

After normalizing the data, the Term Frequency-Inverse Document Frequency (TF-IDF) method was used, which calculates the frequency of a document in the set of documents (corpus). Equation 4.1 shows its formula.

$$tf\_idf_{t,d} = tf_{t,d} \cdot log\ \frac{N}{df_t} \tag{4.1}$$

$tf\_idf_{t,d}$ = Inverse frequency of the term in the document.

$tf_{t,d}$ = Occurency of the frequency of the term t in the document d.

N = Total number of documents.

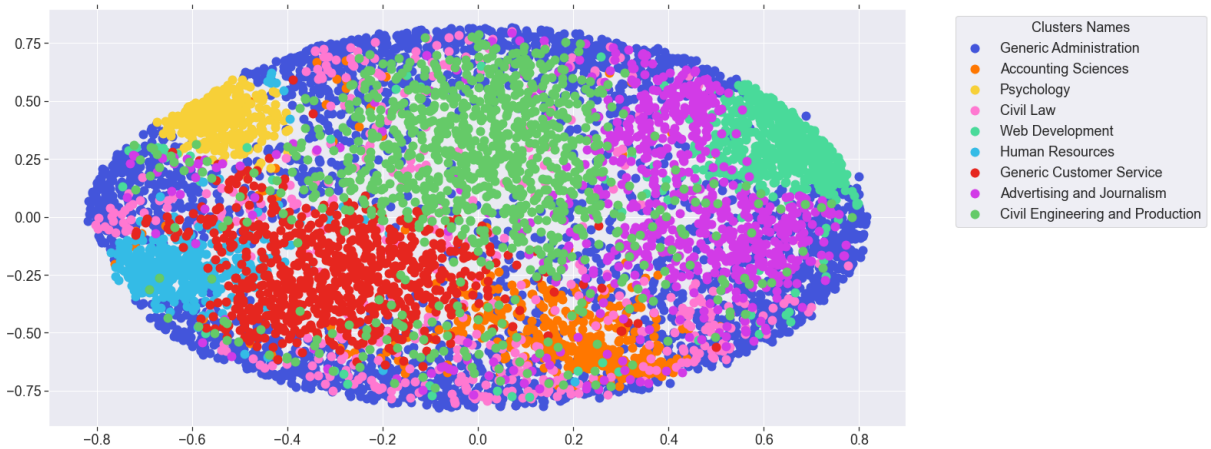
df_t = Number of documents with the term t.

The calculation was done by importing the TfIdfVectorizer method, found in the sklearn module (library also in Python). The result of this method was a numerical set of the frequency of terms in each document. An important point of this stage was the use of bigrams, that is, a bigram is an n-gram for n = 2. This means the composition of terms with two words like “banco dados"(database).

4.4 Clustering

To find the optimal number of clusters, two techniques were used: the Elbow Technique (Elbow Method)(YUAN; YANG, 2019) and the Silhouette (YUAN; YANG, 2019). Both were applied in the range of 2 to 30 K values and they demonstrated that after 9, there were not many variations in the distance of the terms in the set of words. After obtaining this optimum number, the K-means algorithm was used in the clustering, where in Figure 2 the result of the clustering is presented.

Figure 2 – Clustering result of profiles by each cluster.



Source: Made by the author.

After clustering and analyzing the terms that came closest to the centroid, a dictionary of clusters was created with names that made more sense to the words found in each cluster. The suggested names based on course names were: *Generic Administration; Accounting Sciences; Psychology; Civil Law; Web Development; Human Resources; Generic Customer Service; Advertising and Journalism* and *Civil Engineering and Production*. Among the analyzed clusters, there is one in particular that consisted of sets of words present in most of the others, however, without much intensity. This cluster was

called “Generic Administration” and can be seen in dark blue in Figure 2.

Figure 3 – Top 50 words by each cluster.

Source: Made by the author.

In Figure 3, as the words get closer to the center of the cluster, more related among the profiles of that cluster they are and the larger the word size also in the presentation. For example, in the “Civil Law” cluster, the radicals “tributari”, “civil” and “trabalh” are closer to the radical “direit”. This means that resumes of candidates for “Law” job positions have mostly these words highlighted.

4.5 Classification

After the grouped dataset and the respective clusters, the data was divided into 80% training and 20% testing using the module *train_ test_ split* from the sklearn library of the python library to then classify the data. Five classification algorithms were chosen for comparison: *Random Forest* (BICEGO, 2019), *KNN* (NARAYANAN; DJANEYE-BOUNDJOU; KEBEDE, 2016), *SVM* (NARAYANAN; DJANEYE-BOUNDJOU; KEBEDE, 2016), *Decision Tree* (SIVARAM; RAMAR, 2010) and *Naive Bayes* (DURAIRAJ; VIJITHA, 2014).

After the execution of the algorithms, two metrics were obtained, the *Score* where the algorithm with the best classification performance was the *SVM* with 0.9269 according to Table 2 and the metric *Accuracy*, where the algorithm with the best performance was also *SVM* with 94.71%. The accuracy metric was obtained through the *Cross Validation*

Table 2 – Classification Score and Accuracy of Execution Results.

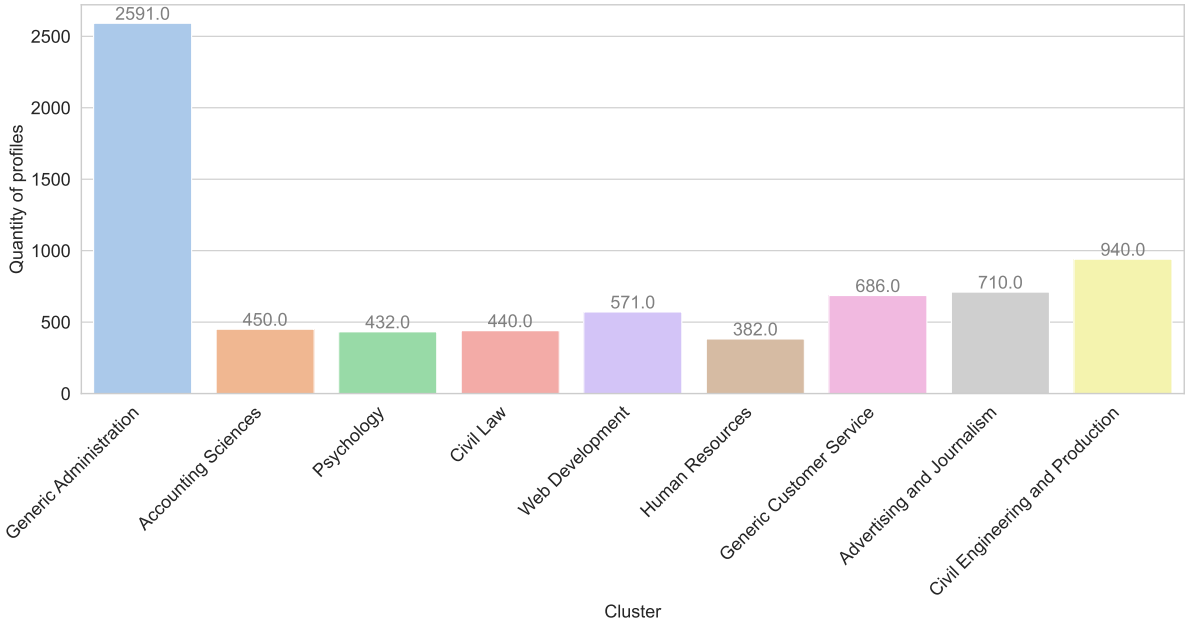
Method	Score	Accuracy (%)
SVM	0.92694	94.71
Random Forest	0.90472	92.29
Decision Tree	0.81725	85.84
KNN	0.75963	77.24
Naive Bayes	0.65741	76.87

Source: Made by the author.

validation method, which consists of assessing the generalization of the model. This metric was also used by (SIVARAM; RAMAR, 2010) and (ROY; CHOWDHARY; BHATIA, 2020) to evaluate the performance of the models. In (ROY; CHOWDHARY; BHATIA, 2020) approach, the best algorithm was also *SVM* with a 78.53% accuracy in their data set. All executions were performed on the same training and test set in each algorithm.

Once the classifier was chosen, in this case, using the SVM algorithm, the profiles were classified. We note in Figure 4 that the generic cluster("Generic Administration") is the largest one with 2,591 CVs. This is due to the low variability of words in the profile or words that are not very relevant to the cluster in question. This 9 clusters covers the majority of the profiles in the private base.

Figure 4 – Plot Quantity Profiles of Clusters.



Source: Made by the author.

5 EXPERIMENTAL RESULTS

5.1 Cosine Similarity

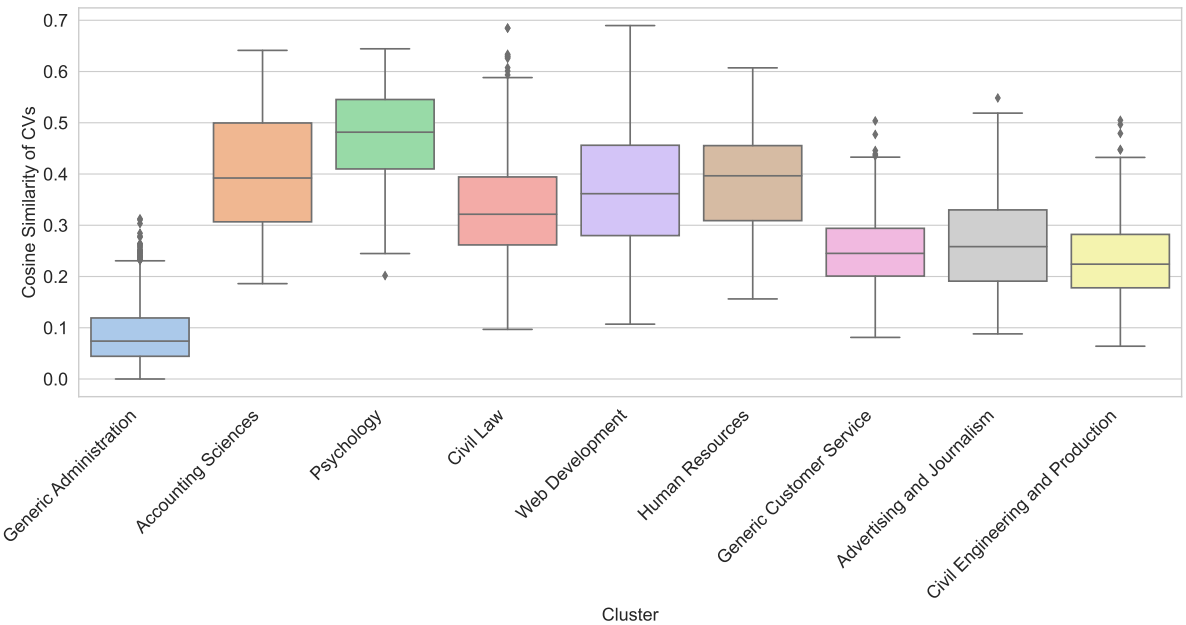
Once the model is trained, and the CVs classified, we need to calculate the distance of each CV to the clusters. Also, it is important to understand the cluster itself, which means, the relevance of the cluster due to vacancies.

The 50 terms (remembering that the bigrams are involved) closest to the cluster's centroid were obtained to calculate the cosine similarity. This calculation measures the similarity of two documents, regardless of the sizes that evaluate the cosine value of the angle between them. Mathematically, each document is projected as a vector into a multidimensional space and then the cosine angle is calculated between two vectors. (KANWAL et al., 2021). Equation 5.1 shows its formula.

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{ab}}{\|\mathbf{a}\|\|\mathbf{b}\|} = \frac{\sum_{i=1}^{n} \mathbf{a}_i \mathbf{b}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{a}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{b}_i)^2}} \tag{5.1}$$

In this case, one of the vectors is the TF-IDF matrix of the 50 words closest to the cluster and the other is the result of this matrix in the vocabulary resulting from the words of the CV itself in each record of the data set.

Figure 5 – Box-plot distribution of the cosine similarity among the clusters.

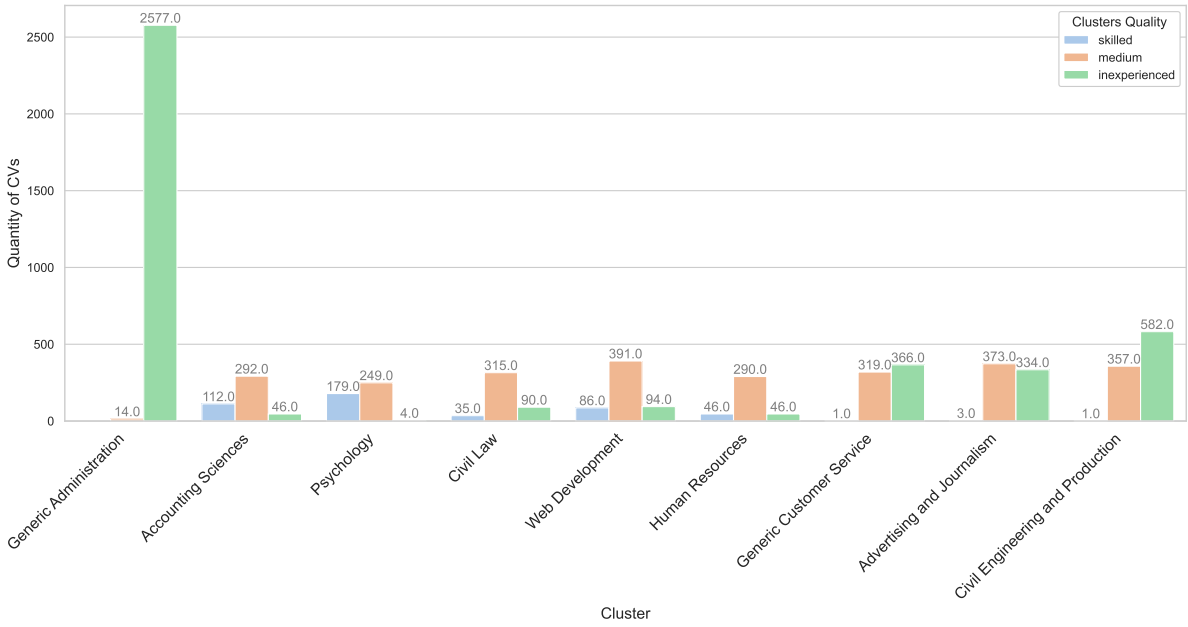


Source: Made by the author.

We can observe in Figure 5 that the cluster with the largest interval is the “Web Development” cluster. This means that this cluster has more adherent profiles and more common words among them. In other words their cosine similarity has the highest levels. In contrast, the “Generic Administration" cluster has the lower levels of cosine similarity.

Once the profiles had the cosine similarity calculated, the cluster similarity was calculated with the words of internship vacancies found on job vacancy sites on the web. These words passed through an intersection, considering only the ones in the defined vocabulary. Using these vacancy words, we created a classification profile quality and the result can be seen in Figure 6.

Figure 6 – Distribution of profiles quality among the clusters.



Source: Made by the author.

Labels were assigned to facilitate the understanding of the similarity calculation. Bellow the following classifications were considered:

1. If the CVs has more than 50% similarity with the cluster it is considered “skilled"
2. If the CVs has between 25% and 50% similarity with the cluster it is considered “medium"
3. If the CVs has less than 25% similarity, it is considered “inexperienced”

For example, a profile with the words "sql sql sql banco dados banco dados redes comp...", with a similarity of 0.276385 has a profile considered “medium" because it does not have yet the most important words for a specific area. This is presented as a feedback to the student.

Table 3 – Sample of Recommendation.

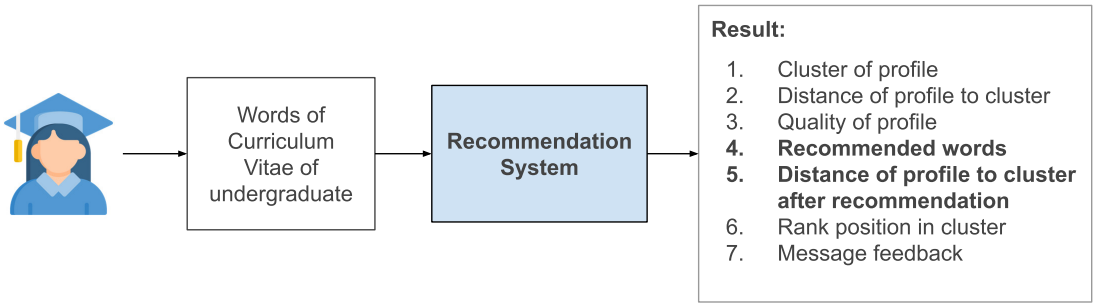
Words in student CV	Recommended Keywords
cinema cinema audiovisual...	portfolio photoshop, spanish, excel word...
calculation accounting hr...	office, selection, department...
accounting accounting nursing...	word office, office, accounting attendance...

Source: Made by the author.

5.2 Recommendation System

Once a candidate has already been classified to a cluster and the similarity of cosine with the cluster was obtained, the result is the recommendation of keywords. The recommendation is based on the most relevant words in the cluster, that is, terms not present in the profile of the current candidate but are present in the profiles of the best candidates. In addition, the recommendation also presents a feedback. Table 3 shows examples of three different profiles with recommendations. Figure 7 shows all the feedback that the model recommends.

Figure 7 – Flow of undergraduate result feedback.



Source: Made by the author.

The Figure 8 provides the complete feedback information for three different CVs with recommendation of 3 terms. Figure 9 provides the complete feedback information for three different CVs with recommendation of 5 terms. Finally, Figure 10 provides the complete feedback information for three different CVs with recommendation of 10 terms.

It is possible to observe in the three records of Figure 10: the keywords extracted from the CV in the column “CV words", the similarity of the cosine calculated with the words about the cluster and the quality and profile. The first record, for example, shows a profile with an “inexperienced” category, which although it has this qualification, this profile is the best within the “Accounting Sciences" cluster with this qualification. Using a recommendation of 10 terms in the first record, as shown in the column “CV words with recommendation", it is possible to notice that if this profile is improved, it can reach a profile with medium quality, increasing the similarity form 0.2498 to 0.4322, according to the column “Cosine similarity of profile after recommendation". The last record in the

Figure 8 – Complete feedback results with 3 words.

CV words	Cosine similarity of profile to cluster	Cluster name	Profile quality	Rank	CV words with recommendation	Cosine similarity of profile after recommendation	Message feedback
labor sales contracts rh office windows informatica service civil law teacher labor sales contracts rh office windows informatica service civil law teacher	0.2484	Civil Law	inexperienced	2nd	labor sales contracts rh office windows informatica service civil law teacher labor sales contracts rh office windows informatica service civil law teacher law law market labor tax contracts	0.3429	You are among the top 10 of the inexperienced profile on your cluster. Check our skills recommendations and improve your profile.
commercial civil production works works autocad autocad excel excel word power point linkedin commercial civil production works works autocad autocad excel excel word power point linkedin market	0.4161	Civil Engineering and Production	medium	9th	commercial civil production works works autocad autocad excel excel word power point linkedin commercial civil works works works autocad autocad excel excel word power point linkedin market english word windows power autocad word	0.4862	You are among the top 10 of the medium profile on your cluster. Check our skills recommendations to improve and have a skilled profile.
works excel works excel word windows power point works excel works excel word windows power point	0.3784	Civil Engineering and Production	medium	29th	works excel works excel word windows power point works excel works excel word windows power point word english production autocad word	0.5007	Your profile is medium. Check our skills recommendations to improve and have a skilled profile.

Source: Made by the author.

Figure 9 – Complete feedback results with 5 words.

CV words	Cosine similarity of profile to cluster	Cluster name	Profile quality	Rank	CV words with recommendation	Cosine similarity of profile after recommendation	Message feedback
spanish rh behavioral behavioral informatics service service service pedagog pedagog spanish rh behavioral behavioral information service service pedagog pedagog	0.2488	Psychology	inexperienced	1st	spanish rh behavioral behavioral informatics service service pedagog pedagog spanish rh behavioral behavioral information service service pedagog pedagog market pedagog documents english informatica pedagog psychology service office pedagog	0.4247	You are among the top 10 of the inexperienced profile on your cluster. Check our skills recommendations and improve your profile.
ml ml marketing contracts contracts personnel department rh rh rh rh rh rh rh human resources human resources human resources human resources human resources excel word selection recruitment recruitment recruitment recruitment recruitment recruitment linkedin linkedin linkedin competences competences competences service attendance ml ml marketing contracts contracts personnel department rh rh rh rh rh human resources human resources human resources human resources excel word selection recruitment recruitment recruitment recruitment recruitment linkedin linkedin linkedin competences competences competences service customer service	0.491	Human Resources	medium	8th	ml ml marketing contracts contracts personnel department rh rh rh rh rh rh rh human resources human resources human resources human resources human resources excel word selection recruitment recruitment recruitment recruitment recruitment linkedin linkedin linkedin competences competences competences service attendance ml ml marketing contracts contracts personnel department rh rh rh rh rh human resources human resources human resources human resources excel word selection selection recruitment recruitment recruitment recruitment linkedin linkedin linkedin competences competences services customer service english resources sales accounting interview human resources financial monitoring spreadsheets	0.5196	You are among the top 10 of the medium profile on your cluster. Check our skills recommendations to improve and have a skilled profile.
scientific initiation congress english exchange spanish spanish event event rh rh rh rh rh rh human resources excel word office recruitment recruitment linkedin indicators scientific initiation congress english exchange spanish spanish event event rh rh rh rh rh rh human resources excel word office recruitment recruitment linkedin indicators	0.2761	Human Resources	medium	264th	scientific initiation congress english exchange spanish spanish event event rh rh rh rh rh rh human resources excel word office recruitment recruitment linkedin indicators scientific initiation congress english exchange spanish spanish event event rh rh rh rh rh rh human resources excel word office recruitment recruitment linkedin indicators human interview financial monitoring resources hr management resources reception	0.3525	Your profile is medium. Check our skills recommendations to improve and have a skilled profile.

Source: Made by the author.

example, on the other hand, shows a profile with a medium quality, which with the new terms, can achieve a “skilled" category, that is, suitable for vacancies, as shown in the “Message feedback" column. It is important to note that the records for recommendation (6739) only involve profiles qualified as “inexperienced” and “medium".

Figure 10 – Complete feedback results with 10 words.

CV words	Cosine similarity of profile to cluster	Cluster name	Profile quality	Rank	CV words with recommendation	Cosine similarity of profile after recommendation	Message feedback
accounting accounting nursing nursing accounting accounting nursing nursing	0.2498	Accounting Sciences	inexperienced	1st	accounting accounting nursing nursing accounting accounting nursing nursing fiscal fiscal marketing point windows commercial law receptionist excel accounting department department personnel linkedin invoice fiscal accounting	0.4322	You are among the top 10 of the inexperienced profile on your cluster. Check our skills recommendations and improve your profile.
accounting secretary receptionist management people human resources accounting accounting secretary receptionist management people human resources accounting	0.2324	Accounting Sciences	inexperienced	17th	accounting secretary receptionist management people human resources accounting accounting secretary receptionist management people human resources accounting fiscal fiscal marketing point windows commercial law excel accounting department department personnel linkedin launch fiscal accounting point point	0.4321	Your profile is inexperienced. Check our skills recommendations and improve your profile.
adobe illustrator photoshop illustrator photoshop design marketing linkedin adobe illustrator photoshop illustrator photoshop design marketing linkedin	0.4652	Advertising and Journalism	medium	9th	adobe illustrator photoshop illustrator photoshop design marketing linkedin adobe illustrator photoshop illustrator photoshop design marketing linkedin graph linkedin behance social media digital content marketing excel premiere attendance networks english campaign media	0.5717	You are among the top 10 of the medium profile on your cluster. Check our skills recommendations to improve and have a skilled profile.

Source: Made by the author.

5.3 Discussion

In Table 4, it is possible to observe the percentage increase in cosine similarity according to the recommendation of keywords for student CVs in each cluster. The calculation of the percentage of increase was obtained from the average similarity before and after the recommendation. On average, all clusters had significant increase when applying recommendations of 3, 5, and 10 keywords. It is important to note that the comparison of these words from the CV together with the recommended keywords were compared and calculated with the 50 keywords closest to the centroid of each cluster, that is, the most important keywords. Thus, the mean of the previous and following clusters of the recommendation was obtained.

Finally, Table 5 shows the Mean Cosine before and after recommendation and the similarity in percentage with 3, 5 and 10 keywords. The results were positive with a minimum increase of 18.83% using only 3 terms and reaching the 50.67% with 10 recommendation words, optimizing the students CV.

Table 4 – Result of recommendation with 3, 5 and 10 words in each cluster.

Cluster name	Number of CVs	Similarity increase by suggesting 3 words (%)	Similarity increase by suggesting 5 words (%)	Similarity increase by suggesting 10 words (%)
Generic Administration	2591	62.41	99.99	146.24
Accounting Sciences	338	12.01	17.32	38.26
Psychology	253	15.78	24.93	33.86
Civil Law	405	27.37	49.62	77.22
Web Development	485	15.24	22.39	36.58
Human Resources	336	9.73	20.07	33.70
Generic Customer Service	685	18.14	28.95	48.63
Advertising and Journalism	707	13.77	21.55	42.36
Civil Engineering and Prod.	939	32.62	57.83	86.32

Source: Made by the author.

Table 5 – Result of general recommendation in clusters.

Recommendation	Mean cosine before recommend.	Mean cosine after recommend.	Similarity increase after keywords recommendation (%)
Recommendation of 3 words	0.289304	0.343768	18.83
Recommendation of 5 words	0.289304	0.379977	31.34
Recommendation of 10 words	0.289304	0.435906	50.67

Source: Made by the author.

6 CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

The grouping of curriculum keywords brought up groups of profiles relevant to the context of the undergraduate students. In addition, it was possible to evaluate the classification of CVs in the groups generated by comparing different types of algorithms, where the algorithm with greater precision was the *SVM* method. It was possible to understand the distance of the profiles of undergraduate students in each cluster and also the distance of this cluster with the vacancies. The use of clustering techniques allowed a reduction and optimization of comparisons of the profile of a candidate with a vacancy.

The feedback for the inexperienced students, in terms of distance from a profile considered "skilled", made students aware about their CV competitiveness for internships in their respective fields. Once the candidate is in a cluster that does not match his course, he is also informed of this profile trend. Another important piece of information is the student rank in a cluster.

The experimental results showed that recommendations improved students CVs similarity (competitiveness within a specific field) from 18.83%, with 3 keywords recommendation, up to 50.67%, with 10 words. It was also possible to provide a feedback message to the candidate, as well as a qualification of his profile and ranking against other candidates.

6.2 Future work

As future work, it is intended to expand the dataset with more analysis to find new profiles of undergraduate CVs. The creation of a synonyms can also be implemented. Also, Deep Learning techniques can be implemented to find a greater number of keywords and skills relevant to the profiles of students who are starting their professional careers. We also intend to create a tool so that candidates can apply the recommendations in practice, optimizing their CVs and being more suitable for vacancies in the professional market.

REFERENCES

ALMALIS, N. D. et al. Fodra - A New Content-Based Job Recommendation Algorithm For Job Seeking And Recruiting. In: IISA 2015 - 6th International Conference on Information, Intelligence, Systems and Applications. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2016. ISBN 9781467393119. 38

BICEGO, M. K-Random Forests: A K-Means Style Algorithm For Random Forest Clustering. In: Proceedings of the International Joint Conference on Neural Networks. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2019. v. 2019-July. ISBN 9781728119854. 34, 44

CAMBRIA, E.; WHITE, B. Jumping NLP Curves: A Review Of Natural Language Processing Research. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2014. 48–57 p. 31

CHALA, S.; FATHI, M. Job Seeker To Vacancy Matching Using Social Network Analysis. In: Proceedings of the IEEE International Conference on Industrial Technology. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2017. p. 1250–1255. ISBN 9781509053209. 39

CHALA, S.; HARRISON, S.; FATHI, M. Knowledge Extraction From Online Vacancies For Effective Job Matching. In: Canadian Conference on Electrical and Computer Engineering. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2017. ISBN 9781509055388. ISSN 08407789. 39

DIABY, M.; VIENNET, E. Taxonomy-Based Job Recommender Systems On Facebook And Linkedin Profiles. In: Proceedings - International Conference on Research Challenges in Information Science. [S.l.]: IEEE Computer Society, 2014. ISBN 9781479923939. ISSN 21511357. 37

DING, Y. et al. A Reciprocal Recommender System For Graduates' Recruitment. In: Proceedings - 2016 8th International Conference on Information Technology in Medicine and Education, ITME 2016. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2017. p. 394–398. ISBN 9781509039050. 27, 37

DIYA, O. Resume Management And Recruitment Workflow System And Method - Google Patents. 2003. Disponível em: <https://patents.google.com/patent/US7711573B1/en>. 27, 30

DURAIRAJ, M.; VIJITHA, C. Educational Data Mining For Prediction Of Student Performance Using Clustering Algorithms. [S.l.], 2014. Disponível em: <www.ijcsit.com>. 33, 34, 38, 44

GUPTA, A.; GARG, D. Applying Data Mining Techniques In Job Recommender System For Considering Candidate Job Preferences. In: Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2014. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2014. p. 1458–1465. ISBN 9781479930791. 37

HARRIS, C. G. Finding The Best Job Applicants For A Job Posting: A Comparison Of Human Resources Search Strategies. In: IEEE International Conference on Data Mining Workshops, ICDMW. [S.l.]: IEEE Computer Society, 2017. v. 2017-November, p. 189–194. ISBN 9781538614808. ISSN 23759259. 36

HONG, W.; ZHENG, S.; WANG, H. Dynamic User Profile-Based Job Recommender System. In: Proceedings of the 8th International Conference on Computer Science and Education, ICCSE 2013. [S.l.: s.n.], 2013. p. 1499–1503. ISBN 9781467344623. 39

KANWAL, S. et al. A Review Of Text-Based Recommendation Systems. In: . [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2021. v. 9, p. 31638–31661. ISSN 21693536. 35, 46

KAUR, B. Data Science: Empowering Business Strategy. In: 2019 IEEE Pune Section International Conference, PuneCon 2019. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2019. ISBN 9781728119243. 30

LIU, W. et al. Optimized Clustering Based On Semantic Similarity Of Components For Short Text. In: 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP). [S.l.]: Institute of Electrical and Electronics Engineers (IEEE), 2020. p. 1–6. 32

MAURYA, A.; TELANG, R. Bayesian Multi-View Models For Member-Job Matching And Personalized Skill Recommendations. In: Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2018. v. 2018-January, p. 1193–1202. ISBN 9781538627143. 38

NARAYANAN, B. N.; DJANEYE-BOUNDJOU, O.; KEBEDE, T. M. Performance Analysis Of Machine Learning And Pattern Recognition Algorithms For Malware Classification. In: Proceedings of the IEEE National Aerospace Electronics Conference, NAECON. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2016. v. 0, p. 338–342. ISBN 9781509034413. ISSN 23792027. 34, 35, 44

NGO, A. Human Resources Internship Report Thesis. [S.l.], 2019. 27, 30

PIERRE, J.; JEANNE, J. M. Exploring The Relationship Between Internship And Employability. [S.l.], 2020. Disponível em: <https://researchspace.ukzn.ac.za/handle/10413/19258>. 27

RAMANATH, R. et al. Towards Deep And Representation Learning For Talent Search At Linkedin. In: International Conference on Information and Knowledge Management, Proceedings. New York, NY, USA: Association for Computing Machinery, 2018. v. 9, n. 18, p. 2253–2262. ISBN 9781450360142. Disponível em: <https://dl.acm.org/doi/10.1145/3269206.3272030>. 27

RODRIGUEZ, L. G.; CHAVEZ, E. P. Feature Selection For Job Matching Application Using Profile Matching Model. In: IEEE 4th International Conference on Computer and Communication Systems (ICCCS). [S.l.]: Institute of Electrical and Electronics Engineers (IEEE), 2019. p. 263–266. 27, 36, 37

ROY, P. K.; CHOWDHARY, S. S.; BHATIA, R. A Machine Learning Approach For Automation Of Resume Recommendation System. In: . [S.l.]: Elsevier B.V., 2020. v. 167, p. 2318–2327. ISSN 18770509. 39, 45

SIVARAM, N.; RAMAR, K. Applicability Of Clustering And Classification Algorithms For Recruitment Data Mining. Kovilpatti, India, 2010. v. 4, n. 5, 975–8887 p. Disponível em: <https://doi.org/10.5120/823-1165>. 32, 33, 38, 44, 45

TURING, A. 1950. Private communication.

WALLER, M. A.; FAWCETT, S. E. Data Science, Predictive Analytics, And Big Data: A Revolution That Will Transform Supply Chain Design And Management. Journal of Business Logistics, Council of Supply Chain Management Professionals, v. 34, n. 2, p. 77–84, 2013. ISSN 21581592. 30

WANG, J. et al. The Design Of University Employment Information Service System Based On Big Data. In: . [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2020. p. 379–382. ISBN 9781728196381. 28, 39

WEI, D.; VARSHNEY, K. R.; WAGMAN, M. Optigrow: People Analytics For Job Transfers. In: Proceedings - 2015 IEEE International Congress on Big Data, BigData Congress 2015. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2015. p. 535–542. ISBN 9781467372787. 35, 36, 37

YUAN, C.; YANG, H. Research On K-Value Selection Method Of K-Means Clustering Algorithm. J—Multidisciplinary Scientific Journal, MDPI AG, v. 2, n. 2, p. 226–235, 6 2019. Disponível em: <https://www.mdpi.com/2571-8800/2/2/16>. 32, 43

ZHENG, S. et al. Job Recommender Systems: A Survey. In: ICCSE 2012 - Proceedings of 2012 7th International Conference on Computer Science and Education. [S.l.: s.n.], 2012. p. 920–924. ISBN 9781467302425. 35