Resume Analysis in Portuguese using Word Embeddings: Development of a Decision Support System for Candidate Selection

Manoel Garcia de Sousa Neto[1], **Filipe Saraiva**[1]

[1]Instituto de Ciências Exatas e Naturais – Universidade Federal do Pará (UFPA)
Caixa Postal 479 – 66.075-110 – Belém, PA – Brazil

manoel.sousa.neto@icen.ufpa.br, saraiva@ufpa.br

Abstract. *This study proposes the use of word embeddings to analyze and compare resumes in Portuguese with job requirements. It presents the development of a decision support system that effectively identifies the suitability of candidates based on their professional experience. The study employs different word embedding models, such as Word2Vec, Wang2Vec, FastText, and GloVe, to generate numerical representations of words in resumes and job descriptions to benchmark these models in a Portuguese-language context. The research aims to assist resume analysis and enhance the accuracy of candidate selection processes.*

Resumo. *Este estudo propõe o uso de word embeddings para analisar e comparar currículos em português com descrições para vagas de emprego. Apresenta o desenvolvimento de um sistema de suporte à decisão que identifica a adequação dos candidatos com base em sua experiência profissional. O estudo utiliza diferentes modelos de word embeddings, como Word2Vec, Wang2Vec, FastText e GloVe, para gerar representações numéricas de palavras em currículos e descrições de empregos, a fim de avaliar o desempenho desses modelos no contexto da língua portuguesa. A pesquisa tem como objetivo auxiliar na análise de currículos e aprimorar a precisão dos processos de seleção de candidatos.*

1. Introduction

In the last two decades, with the advent of the Internet and the vast increase in the number of users, the current situation of the Internet, according to Ivancevich (2008), has brought about a revolution in recruitment processes. This is due to the ability to widely propagate recruitment processes among a large number of active Internet users, enabling people from diverse locations to apply for job vacancies through web-based platforms. In many cases, this process is guided by questionnaires or pre-defined attributes that include general characteristics in exams or profiles on websites and may also submit detailed resumes that are directed towards recruiters who analyze them to determine the suitability of candidates for the job[Medeiros 2017]. The goal of organizations is to always hire employees who are a perfect fit for the job. However, if a mistake is made in the selection process for a new employee, it can have serious financial consequences for the organization. This puts a great deal of pressure on the individuals responsible for making the hiring decision, as they must carefully evaluate multiple applicants to find the most suitable candidate.

To begin the recruitment process, organizations first announce an open job position using a variety of channels such as websites, newspapers, and other media. Prospective candidates who are interested in the position can then apply by either creating a profile using a designated online form or by uploading their resumes through the organization's website. The received applications are then subjected to a thorough evaluation to identify the most qualified candidates, who are then chosen for an interview[Devi and Banu 2014].

With the goal of applying to the context of resume analysis, this work aims to contribute to the use of term representations using word embeddings technology to identify information contained in resumes and compare it with the job requirements described in the company's job offer, without the need for specific text structures to perform analyses. This allows word embeddings to be used in texts with variations in writing and still be effective in performing classifications [Souza et al. 2020].

Therefore, this paper is justified by: 1) comparing four important word embedding models and chosing one model that expedites a critical step in most job selection processes, which is resume analysis [Oliveira and Macêdo 2020]; 2) contributing to another use of word embeddings in the Portuguese language in a practical task of textual classification; and 3) cooperating with the development of decision support systems in the context of e-recruiting.

Thus, this paper aims to contribute to the study and development of a decision support system that will use word embeddings to analyze existing text in Portuguese-language resumes, compare them with job vacancies available on the platform, and map the meanings present in the resumes involving user characteristics based in professional experience.

This article is structured into five main sections to comprehensively address the research objectives. The introduction provides an overview of recruitment processes and emphasizes the need for accurate candidate selection. The related works section discusses previous studies on NLP techniques in e-recruiting and word embeddings applied to Portuguese language tasks. The materials and methods section outlines the research methodology and tools used. The results and discussion section presents the findings from resume matching and algorithm evaluations. Finally, the conclusion summarizes the contributions of the research and discusses future directions.

2. Related Works

In recent years, the increasing use of e-recruiting has led to the development of various techniques and tools for automating and improving the candidate selection process. Natural Language Processing (NLP) is one such technique that has gained popularity due to its ability to analyze large amounts of unstructured text data, such as resumes and job descriptions, and extract relevant information for decision-making.

Several studies have explored the use of NLP for candidate classification in e-recruiting. These studies typically involve the use of machine learning algorithms, such as support vector machines (SVMs), decision trees, and neural networks, to analyze candidate resumes and job descriptions and make predictions about their suitability for a particular job.

In this chapter, we will present studies that aimed to classify candidates in e-

recruiting processes, primarily based on the use of NLP, as well as studies that perform classifications based on elements such as academic background or professional experience of the candidates. In order to colaborate to Portuguese language word embedding scenario, will also be shown studies that compare and evaluate the use of portuguese language word embeddings into specific domains.

In the article by [Najjar et al. 2021], a decision support system is applied in the context of a web application that ranks candidates according to the semantic similarity of resumes and job descriptions, using word embeddings such as Word2Vec to vectorize the analyzed texts and use the semantic representation of words in vectors to then use cosine similarity to compare the texts belonging to the resumes and job offers, placing resumes that had higher semantic similarity as better placed and selecting only the top three. This study has many elements that show that the use of word embeddings for the purpose of this article results as precise tool, compared to the studies conducted by [Mohamed et al. 2018], [Sivaramakrishnan et al. 2018], and [Roy et al. 2020] that are also mentioned in this related works section, showing higher accuracy in the results compared to the studies mentioned earlier.

In the study by [Mohamed et al. 2018], the Smart Applicant Ranker resume recommendation system is proposed, which acts to apply OWL ontology and natural language processing to identify and classify candidates' resumes to job requirements models, using semantic similarity in screening resumes. However, it does not present itself as a comprehensive solution in many contexts since it was designed to operate only in the scope of recruiting candidates for the IT area and shows that it needs restructuring to be applied in other ways.

In the study by [Sivaramakrishnan et al. 2018], an employment recommendation system is shown that is primarily based on classifying resumes by similarity using VSM (Vector Space Model) methods, where classifiers by term frequency and cosine similarity are used to recommend the documents with the highest similarity. Due to the fact that classifiers by term frequency do not orient themselves by the positioning of words in the texts and focus more on determining the importance of recurring terms, there are difficulties in highlighting specific key terms in the text that are not as recurrent.

The study by [Roy et al. 2020] addresses a candidate resume recommendation system based on machine learning that seeks to indicate the resume closest to the job description. In the processing of this recommendation system, classification models such as Random Forest, Naive Bayes, Logistic Regression, and linear SVM are applied to summarized texts using term frequency, and after this processing, they are ranked based on cosine similarity to choose the best resumes. Due to summarization, many relevant data are lost, which ends up altering the analysis results, affecting the accuracy of the proposed classification system.

The works of Hartmann et al. (2017) and Rodrigues et al. (2016) establish a foundation for studies on word embeddings applied to the Portuguese language, taking initial steps to explore the differences between word embedding models and comparing them in tasks of semantic, syntactic, and morphological similarities in the overall context of the Portuguese language. Thus, due to the specific characteristics observed in the use of word embeddings in the Portuguese language, many studies utilizing them in specific

domains have been conducted as Martins and Silva(2021) applying word embeddings in law area, Bonadia and Barreto(2019) in sentiment analysis and Pastro(2018) using for traffic purposes.

In the study conducted by Martins and Silva (2021), a word embeddings approach to the legal domain context is presented. The study explores the application of word embeddings models such as Word2Vec, ELMO, FastText, GloVe, and Wang2Vec in document classification tasks, specifically for migrating physical processes to electronic systems and automating the routing of legal demands. The authors suggest that employing these techniques can expedite processes within the Brazilian Justice System.

The work by Bonadia and Barreto (2019) proposes the analysis of the use of word embeddings in the Portuguese language for sentiment analysis. They utilize datasets composed of movie reviews to determine the accuracy of methods such as bag of words and word embeddings, and to perform comparisons between these methods. Additionally, they compare the results applied in the context of the Portuguese language with classifiers operating in English language contexts.

The work by Pastro (2018) discusses the applicability of models such as Word2Vec, Doc2Vec, and FastText in detecting traffic events reported on Twitter. The study involves a large dataset to identify traffic incidents and compares the accuracy of the models in semantically filtering and retrieving the required texts.

Thus, this work aims to explore the use of word embeddings in Portuguese for the e-recruitment scope, acting as an NLP screening tool and contributing to Portuguese language NLP in applied tasks by providing a solid comparison between word embedding models.

3. Materials and Methods

This article is an applied research of quantitative nature, which is developed through methods and tools capable of handling data collected during the research and extracting information from the analysis via natural language processing mechanisms.

The quantitative factor is about the rankings made between the applicants ordered by the most suitable for the job to the less suitable, using the results of semantic similarity detected and comparing word embedding results with human recruiter results.

This study has characteristics of an exploratory research, as it seeks to catalog phenomena [Oliveira 2011] occurring in the classificatory nature of the word embeddings technology, used in a real context and directed towards the specificities of the technology's functioning in the context of the Portuguese language.

3.1. Development Materials

Among the tools used, due to the applied nature of this work, software focused on system development, natural language processing, and word embeddings available in the repository of word embeddings of the Núcleo Interinstitucional de Linguística Computacional [NILC-USP 2017], which constitutes a large amount of word embeddings in portuguese that is being used to assist in comparisons between resumes and anonymous job descriptions provided in a Brazilian employment platform Talentos Carreira RH.

Tools such as the Python programming language version 3.10 was used due to compatibility with the computational intelligence libraries being used, such as Gensim for processing related to word embeddings, Natural Language Toolkit (NLTK) for pre-processing step.

3.2. Word Embeddings

Word embeddings are a type of natural language processing (NLP) technique that represents words as vectors of numerical values. These values are generated by a mathematical model that takes into account the context in which the words appear in a corpus of text. The purpose of word embeddings is to provide a way to represent words in a way that a machine can understand and process, enabling the machine to better understand the meaning and relationships between words in a given language[Souza et al. 2020].

By using word embeddings, machine learning models can better identify and classify text, and can even generate new text that has similar meaning and tone to existing text. Word embeddings have become an important tool in a wide range of NLP applications, such as sentiment analysis, text classification [Zhai et al.2016].

In this study, the types of word embedding architectures that are being used are Word2Vec, Wang2Vec, FastText and GloVe.

3.3. Word Embedding Models

3.3.1. Word2Vec

Created by Mikolov et al. (2013), Word2Vec is a popular unsupervised learning algorithm used for natural language processing. It is used to generate a numerical representation, or embedding, of words in a large corpus of text. Word2Vec works by training a neural network on a large corpus of text to predict surrounding words given a center word. The algorithm uses two main models: Continuous Bag-of-Words (CBOW) and Skip-Gram.

CBOW predicts a center word based on surrounding words, while Skip-Gram predicts surrounding words given a center word. In both models, the neural network is trained to minimize the difference between the predicted and actual surrounding words, which leads to the formation of word embeddings.

The Skip-Gram model uses a neural network with one hidden layer and is trained to minimize the difference between the predicted context words and the actual context words. During the training process, the model learns the relationships between the center word and the context words. In summary, the Skip-Gram model in Word2Vec takes a center word as input and predicts surrounding context words, and the resulting word embeddings capture the semantic and syntactic relationships between words[Mikolov et al.2013].

The word embeddings generated by Word2Vec capture semantic and syntactic relationships between words, allowing them to be used in various NLP tasks, such as text classification, clustering, and similarity comparison[Mikolov et al.2013].

3.3.2. Wang2Vec

Presented by [Ling et al. 2015], Wang2Vec is based on many features present in Word2Vec, but proposes the use of structured skip-grams, which, according to the authors, is capable of improving aspects of the analysis of the syntactic behavior of words by using a different set of parameters for each context word, varying according to its position in relation to the target word.

This represents the biggest difference compared to the skip-gram proposed in [Mikolov et al. 2013]. Wang2Vec is pointed out as the most accurate architecture for syntactic analysis activities in the application for the Portuguese language [Hartmann et al. 2017].

3.3.3. FastText

Introduced by [Bojanowski et al. 2017] and based in Word2Vec, FastText works by training a supervised deep neural network to predict the likelihood of a word given its surrounding context. Unlike other word embedding algorithms, such as Word2Vec, FastText takes into account subword information in addition to whole word information. This allows FastText to generate embeddings for rare or out-of-vocabulary words by combining the embeddings of its subwords. FastText also allows for text classification by adding a simple, fully connected layer on top of the word embeddings. This layer is trained to predict the class labels for a given input text.

In summary, FastText is a word embedding algorithm that generates numerical representations of words and subwords in a large corpus of text, and can also be used for text classification. Its unique approach of taking into account subword information allows it to generate embeddings for rare or out-of-vocabulary words.

3.3.4. GloVe

Presented by Pennington et al.(2014), GloVe is a word embedding algorithm that trains a regression model to predict the co-occurrence counts of words in a corpus, leading to the generation of numerical representations of words, known as word embeddings, that capture both semantic and magnitude relationships between words.

GloVe works by training a regression model to predict the co-occurrence counts of words in a corpus. The co-occurrence count of two words refers to the number of times they appear together in a context window. The regression model is trained to minimize the difference between the predicted and actual co-occurrence counts, which leads to the formation of word embeddings[Pennington et al. 2014].

GloVe's embeddings capture not only semantic relationships between words but also the magnitude of the relationships. This is achieved by using a weighted combination of the global word-word co-occurrence information and local word-context information.

3.4. Implementation Steps

3.4.1. Acquisition and Pre-processing

The initial steps consist of acquiring the documents that will be used as a base in the classification process, provided by the Talentos Carreira RH platform, documents related to job requirements and resumes of anonymous participants in the process. After obtaining the corpus of terms belonging to the resumes job experience sections and job descriptions, the terms are pre-processed using the NLTK and Spacy libraries [Anchiêta et al. 2021] to fulfill pre-processing steps related to:

- Sentence segmentation;
- Removal of irrelevant words for subsequent text analysis such as the use of prepositions, definite and indefinite articles, and words that do not add value to the reading;
- Performing Pos-Tagging, which is labeling of textual elements to identify the grammatical structure of the analyzed corpus;
- Text normalization, removing punctuation, converting all uppercase letters to lowercase, removing images, links, and repeated letters in the process;
- Performing term lemmatization to facilitate analysis of morphologies that may vary in the writing of participants in order to reduce noise in the comparison between texts.

3.4.2. Matching between resumes and job descriptions

After the preprocessing is done, the texts are forwarded for analysis using the Gensim library with the aim of vectorizing the sentences, based on the pre-trained word embedding models previously mentioned. After generating the vectors related to the input text, the resume models are compared to the pre-processed and vectorized model of the job requirements description.

The comparison to measure the semantic similarity between the resumes is calculated through the cosine similarity method, which besides being widely used for measuring semantic similarity in natural language processing, has also been shown by [Rodrigues et al. 2016] and [Hartmann et al. 2017] to be efficient in measuring semantic similarity in word embeddings adapted to the Portuguese language.

The cosine similarity serves to measure the similarity between vectors. In the context of this article, the vectors will represent the documents to be compared in the cosine angle result, where the closer this angle is to the value 1.0, the greater the similarity found in the comparison of the vectors [Sousa, 2018], as shown in Figure 1, where C are the vectors of the resume model, V are the vectors of the job requirements, and Ø is the angle between the vectors of the compared resumes job experience. After the comparison is done by cosine similarity, the similarity results are ordered by nominal value from highest to lowest among the resumes.

$$Similarity(\vec{C}, \vec{V}) = \cos \emptyset = \frac{C \cdot V}{\|C\|\|V\|} = \frac{\sum_{i=1}^{n} C_i \times V_i}{\sqrt{\sum_{i=1}^{n} C_i^2} \times \sqrt{\sum_{i=1}^{n} V_i^2}} \quad (1)$$

Figure 1. Cosine Similarity calculation.

3.4.3. Extraction and Classification of Results

After the matching stage, the proportionality in relation to the scales used in the classification of the results is performed, since the semantic similarity results vary proportionally in nominal similarity found in comparisons between different areas. Then, according to the proportion of the results, the resumes will be classified from the highest aptitude to lowest in relation to the comparison with the job description offered.

4. Results and Discussion

The process begins with the matching of resumes, utilizing cosine similarity as a measure. This technique generates values ranging from 0 to 1, which represent the degree of semantic similarity. Values closer to 1 indicate a higher proximity in terms of semantic similarity, while values closer to 0 indicate a lower degree of similarity. Once the semantic similarity data is obtained, the resumes are ranked based on their suitability for the job description.

To mimic the behavior of a human recruiter, a comparison was conducted between the rankings generated by different algorithms and the rankings determined through manual evaluation. This comparison aimed to assess the effectiveness of the algorithms in mimicking the decision-making process of a human recruiter.

The comparisons were performed across four distinct job positions, each with groups of ten applicants. For each job position, the analysis focused specifically on the section of professional experience contained within the resumes. By examining the rankings generated by both the algorithms and the manual evaluation, insights were gained regarding the alignment between the automated approaches and the subjective judgments made by a human evaluator.

Table 1 presents similarity values results for a matching between a job description for engineering technical assistant and the extracted work experience, showing values that are closer to each other in terms of applicants ranking but very different in numerical results, while Word2Vec and GloVe performed well in distinguishing strong resumes from weak resumes, FastText and Wang2Vec showed values with less disparity in relation to less qualified candidate resumes.

Table 2 presents the texts used for the position of Technical Engineering Assistant after pre-processing. We can observe that resumes C9 and C1, which have a closer proximity to the job description, contain content that is more relevant to topics related to civil engineering assistant. On the other hand, resume C5 showcases work experience that aligns more with the fields of computing and transportation, thus making it further from the job description due to fewer semantic similarities.

The evaluation process helps to determine the reliability and accuracy of the algorithms in effectively ranking the resumes based on their compatibility with the job requirements. It also provides valuable insights into the strengths and limitations of each algorithm in capturing the nuances of semantic similarity and its relevance to job descriptions.

Table 1. Cosine similarity matching results for engineering technical assistant job.

Engineering Technical Assistant							
Word2Vec		FastText		GloVe		Wang2Vec	
CV	Similarity Value	CV	Similarity Value	CV	Similarity Value	CV	Similarity Value
C9	0.854	C1	0.949	C1	0.929	C1	0.946
C1	0.846	C9	0.943	C9	0.915	C9	0.943
C8	0.791	C8	0.911	C3	0.894	C8	0.910
C2	0.786	C2	0.909	C8	0.884	C2	0.90093
C3	0.775	C10	0.900	C10	0.866	C3	0.90014
C4	0.734	C3	0.898	C2	0.858	C6	0.8927
C10	0.720	C6	0.887	C6	0.838	C10	0.892
C6	0.717	C4	0.882	C4	0.835	C4	0.879
C7	0.568	C7	0.750	C7	0.591	C7	0.755
C5	0.362	C5	0.689	C5	0.427	C5	0.627

Table 2. Pre-processed job description and resumes C9, C1 and C5.

Job Description	"cargo assistente técnico em engenharia segmento construtora necessário formação em engenharia civil cnh disponibilidade para viagens domínio em autocad necessário ter domínio em planilhas excel principais atividades atuar com acompanhamento fiscalização e gestão de obras de acordo com o contrato licitatório elaborar planilhas medições diários de obra e relatórios fotográficos levantamento dos orçamentos cotações e compras essencial perfil técnico bom relacionamento interpessoal boa comunicação proatividade trabalho em equipe"
C9	"técnico operacional especial nivel superior strans teresina limpserv eireli contratual equipe financeiro fiscal consultor equipamentos fiscalização eletrônica radares sinalização viária teresina elaborar acompanhar medições contratuais analista licitação analista processos assistente engenharia strans teresina certare engenharia consultoria ltda fiscalização técnica financeira consultoria contratos equipamentos fiscalização eletrônica radares teresina elaborar acompanhar medições analista processos elaboração projetos sinalização viária orçamento projetos auxiliar administrativo cld construtora laços detetores eletrônica ltda prestadora serviço strans atender orientar público auxiliar fechamento frequência pessoal verificar cumprimento normas administrativas operacionais inspeção controle/roteirização frotas veículares elaborar fazer controle planilhas banco dados relacionados manutenção semafórica estações terminais teresina-piauí controlar entrada saída material estagiario techportas tech montagem comércio portas ltda acompanhamento rotina produção material"
C1	auxiliar técnico volume construções e participações realizar levantamento de quantidade de materiais pelo projeto e acompanhar diretamente com os fornecedores cronograma para atender prazo estabelecido de entrega elaborar solicitação de compra de materiais levantamento de quantitativo de cabos de/para pelo projeto elétrico em autocad planilhas de controle de serviços controle de materiais e andamento de obra supervisão de serviços de instalação e manutenção em eletricidade auxiliar os engenheiros em todos os processos da obra construção civil mecânica e elétrica supervisão de execução de serviços nos sistemas de alarme e detecção de incêndio automação e cabeamento estruturado apoio técnico na operação e manutenção das instalações durante e após os jogos olímpicos e paralímpicos rio 2016 auxiliar técnico controle de materiais e andamento de obra auxiliar os engenheiros de instalações em todos os processos da obra apoio técnico durante a realização do evento teste aquece rio para os jogos olímpicos e paralímpicos rio 2016
C5	estagiária sdu-leste ótima técnico informática procuradoria geral município pgm responsável suporte técnico manutenção equipamentos informática verificação problemas erros hardware software encarregado instalação configuração redes verificação solução vulnerabilidades segurança implantação módulo inventário equipamentos ti fiscalização contratos ti estágio superintendência municipal transportes trânsito strans encarregado alimentação sistema cadastro táxi moto táxi ônibus coletivos transporte alternativo gerência licenciamento concessão glc instalação compartilhamento impressoras computadores

Therefore, since the system aims to cover various types of recruitment processes and handle screenings in different job descriptions regarding candidates with professional experiences from different fields, the use of the extensive corpus of word embeddings provided by NILC-USP is justified. The rankings derived from the analysis of word embedding models were compared to the human ranking, as shown in Table 3, Table 4(Table 4 only presents 9 candidates because C10 does not have any professional experience), Table 5, and Table 6. The values represented correspond to the positions of the candidatesin the ranking, while the values from C1 to C10 are the identifiers of the candidates.

Table 3. Cosine similarity matching results for engineering technical assistant job.

Engineering Technical Assistant					
	Human Ranking	Word2Vec	FastText	GLOVE	Wang2Vec
c1	2	2	1	1	1
c2	3	4	4	6	4
c3	4	5	6	3	5
c4	5	6	8	8	8
c5	10	10	10	10	10
c6	8	8	7	7	6
c7	9	9	9	9	9
c8	7	3	3	4	3
c9	1	1	2	2	2
c10	6	7	5	5	6

Table 4. Cosine similarity matching results for pedagogical coordinator.

Pedagogical Coordinator					
	Human Ranking	Word2Vec	FastText	GLOVE	Wang2Vec
c1	1	2	1	1	1
c2	5	8	8	8	8
c3	6	4	5	5	3
c4	2	1	2	2	4
c5	3	5	4	4	3
c6	4	3	3	3	5
c7	7	6	6	6	6
c8	9	9	9	9	9
c9	8	7	7	7	7

Table 5. Cosine similarity matching results for administrative manager.

Administrative Manager					
	Human Ranking	Word2Vec	FastText	GLOVE	Wang2Vec
c1	1	4	6	8	6
c2	2	3	3	4	3
c3	5	6	7	5	7
c4	4	2	2	2	2
c5	3	1	1	1	1
c6	9	9	9	10	9
c7	6	7	4	6	5
c8	7	5	5	3	4
c9	8	8	8	7	8
c10	10	10	10	9	10

Table 6. Cosine similarity matching results for journalist.

Journalist					
	Human Ranking	Word2Vec	FastText	GLOVE	Wang2Vec
c1	6	6	5	5	5
c2	4	5	8	8	8
c3	1	2	2	2	2
c4	2	3	4	3	3
c5	7	7	6	7	7
c6	10	10	10	10	10
c7	8	8	7	6	6
c8	5	1	1	1	1
c9	3	4	3	4	4
c10	9	9	9	9	9

In order to compare which model is closest to human results was used the Spearman's rank correlation[Spearman 1904] coefficient between the human ranking and each model's ranking.The model with the highest coefficient indicates a stronger relationship and thus a closer resemblance to the human ranking in terms of order. To represent the scores of each model the Spearman's coefficient ranges from -1 to 1, where 1 indicates a perfect mono-tonic relationship, 0 indicates no monotonic relationship, and -1 indicates a perfect inverse monotonic relationship. Figure 1 shows that the model's rankings share a close resemblance while compared to the human ranking, having ratings scoring above 0.7 which means an accuracy above 70% in most of the cases analyzed.

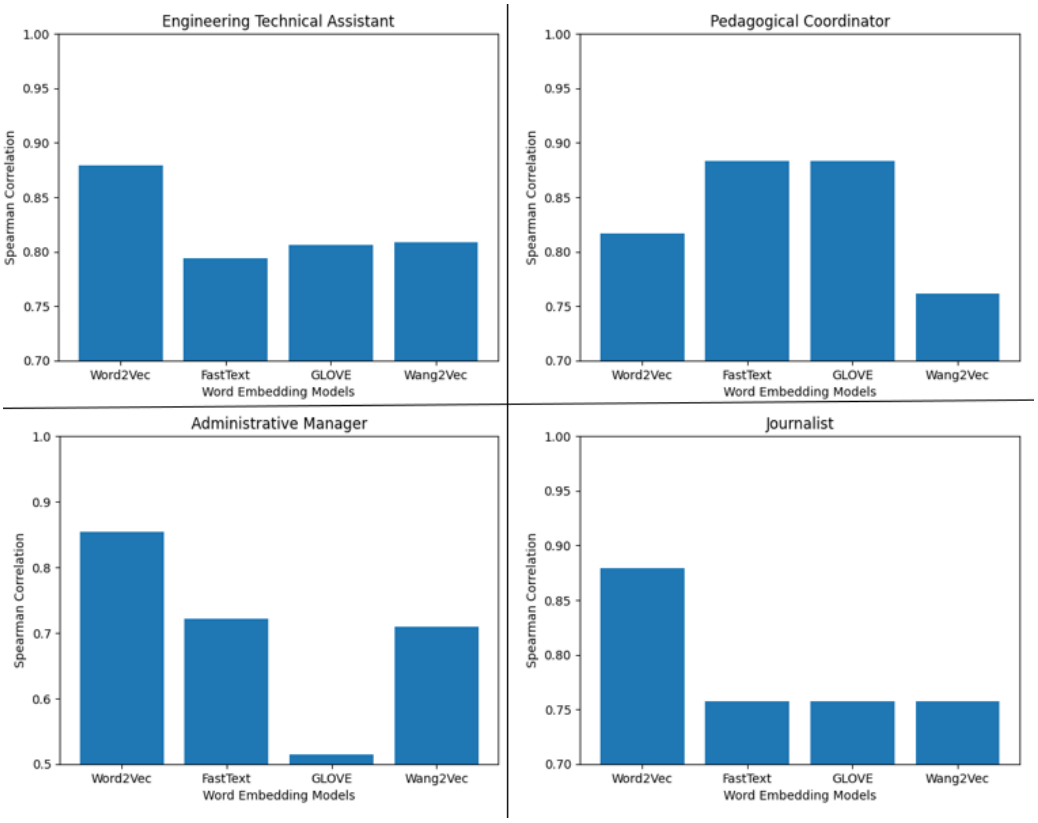

Figure 2. Comparison between Word2Vec, FastText, GLOVE and Wang2Vec using Spearman's coefficient.

Despite the proximity of the model's results in terms of accuracy, only Word2Vec has scored more than 80% accuracy in every test, achieving more solid results in comparison to the rankings generated by other models matchings and also top scoring in 3 of 4 cases of e-recruitment tasks. According to the testing trials focused in recruiting tasks matching job descriptions and applicants job experience, the accuracy ratings for each model are: 85% for Word2Vec, 79% for FastText, 74% for GLOVE and 76% for Wang2Vec.

In conclusion, the experimental results indicate that Word2Vec consistently outperformed other models, achieving an accuracy rate of 85% in matching job descriptions and applicants' job experience. Its superior performance and stability make it a reliable choice for e-recruitment tasks. Although FastText, GLOVE, and Wang2Vec also exhibited relatively high accuracy rates, Word2Vec demonstrated more robust results across multiple tests. These findings highlight the effectiveness of Word2Vec in improving the accuracy and reliability of candidate screening processes in e-recruitment.

5. Conclusions

Overall, the research in this article aims to make valuable contributions to e-recruitment and decision support systems. By using word processing and NLP techniques, organizations can optimize their resume screening process, improve candidate selection, and ultimately make employment decisions recruiting better. The results of this study have the potential to simplify the recruitment process, reduce costs and increase the accuracy of candidate screening.

Further research and development in this area can explore additional word embedding architectures with Large Language Models approaches like BERTimbau Portuguese embeddings [Souza et al. 2020], investigate the impact of different preprocessing techniques, and refine the decision support system to address specific

industry requirements and languagenuances. The continuous advancement of NLP and machine learning technologies offers great potential for optimizing e-recruiting processes and improving the overall efficiency and effectiveness of candidate selection.

As future work, investigating the applicability of fuzzy logic and neural networks into NLP decision support systems to enhance candidate screening accuracy are propitious ways to explore more about Portuguese language nuances and develop more accurate decision systems.

6. Acknowledgement

The authors acknowledge the support of the Brazilian government through the National Council for Scientific and Technological Development (CNPq), National Financer of Studies and Projects (FINEP), Foundation for Research Support of the State of Piauí(FAPEPI) and Carreira RH's support and data provided.

References

Anchiêta, Rafael; Neto, Francisco A.R.; Marinho, Jeziel C.; Moura, Raimundo(2021). PLN: Das Técnicas Tradicionais aos Modelos de Deep Learning. SCB OPEN LIB, Cap.1. p 3-5.

Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, Tomas. Enriching word vectors with subword information(2017). Transactions of the Association for Computational Linguistics, 5:135–146.

Bonadia, Graziella C. Barreto, Gilmar(2019).Análise de Sentimentos em comentários na língua portuguesa: uma comparação de métodos. Simpósio Brasileiro de Automação Inteligente, SBAI 2019, Brazil.

Devi, Renuka B.;Banu, Vijaya(2014). Introduction to Recruitment. SSRG International Journal of Economics and Management Studies 1.2, p.5-8.

Hartmann, Nathan et al(2017). Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks.arXiv preprint arXiv:1708.06025.

Ivancevich, J. M(2008). Gestão de recursos humanos. São Paulo: McGraw-Hill.

Ling, W.; Dyer, C.; Black, A. W.;Trancoso, I(2015). Two/too simple adaptations of word2vec for syntax problems. Em Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p. 1299–1304.

Martins, V. S. Silva, C. D(2021).Text Classification in Law Area: a Systematic Review.Symposium on Knowledge Discovery, Mining and Learning, KDMILE 2021.

Medeiros, Morgana (2017).F.Recrutamento E Seleção De Pessoas: Métodos E Técnicas Que Podem Ser Utilizados Por Profissionais De Recursos Humanos.2017; Monografia; (Aperfeiçoamento/Especialização em Gestão de Pessoas) - Universidade do Sul de Santa Catarina.

Mikolov, Tomas et al(2013). Efficient Estimation of Word Representations in Vector Space. ArXiv13013781.

Mohamed, Ashif; Bagawathinathan, Wickram; Iqbal, Usama; Shamrath, Shahik; Jayakody, Anuradha(2018). Smart Talents Recruiter - Resume Ranking and Recommendation System. 2018 IEEE International Conference on Information and Automation for Sustainability (ICIAfS). Sri Lanka.

Najjar,Arwa; Amro,Belal; Macedo, Mario(2021). An Intelligent Decision Support System For Recruitment: Resumes Screening and Applicants Ranking. Computers & Applied Sciences Complete, Informatica (Ljubljana), Vol.45 (4).

Núcleo Interinstitucional de Linguística Computacional (NILC)(2017). Repositório de Word Embeddings do NILC. Universidade de São Paulo. Available in: Http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc.

Oliveira, Maxwell Ferreira(2011). Metodologia científica: um manual para a realização de pesquisas em Administração. Catalão: UFG.

Oliveira, Ikaro Ramon Vidal De; Macêdo, Maria Erilúcia Cruz(2020). O Papel Determinante do Currículo no Processo de Recrutamento e Seleção / The Role of Curriculum in the Recruitment and Selection Process. ID on line. Revista de psicologia, [S.l.], v. 14, n. 49, p. 212-228.

Pastro, Jonata Teixeira(2018). Detecção de tweets sobre Eventos de Trânsito usando Word Embedding. Monografia(Bacharelado em Ciência da Computação), Instituto de Informática, Universidade Federal do Rio Grande do Sul, Brazil.

Pennington, Jeffrey; Socher, Richard; Manning(2014), Christopher.GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empiricial Methods in Natural Language Processing.

Rodrigues, J. et al(2016). Lx-dsemvectors: Distributional semantics models for portuguese. In: Computational Processing of the Portuguese Language. Cham: Springer International Publishing, p. 259–270. ISBN 978-3-319-41552-9.

Roy, Pradeep Kumar; Chowdhary, Sarabjeet Singh; Bathia, Rocky(2020). A Machine Learning approach for automation of Resume Recommendation system. Procedia Computer Science, vol. 167. p 2318-2327.

Sivaramakrishnan, N; Subramaniyaswamy. V, Arunkumara, S(2018). Validating effective resume based on employer's interest with recommendation system. Int. J. Pure Appl. Math., volume 119.

Sousa, Priscila Sad(2018). Estimando Similaridade Entre Entidades Quando Apenas Seus Nomes Estão Disponíveis. Dissertação(Mestrado em Ciência da Computação)-Departamento de Computação - DECOM, Universidade Federal de Ouro Preto, Brazil.

Souza, Luiz Fernando; Gonçalves, Alexandre Leopoldo; Souza, João Artur(2020). Utilização Prática de Word Embeddings Aplicada à Classificação de Texto. X Congreso Internacional de Conocimiento e Innovación. Panama.

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. Am. J. Psychol. 15, 72–101.

Zhai, M.; Tan, J.; Choi, J.D(2016). Intrinsic and Extrinsic Evaluations of Word Embeddings. In Proceedings of the Thirtieth AAAIConference on Artificial Intelligence (AAAI'16), Phoenix, AZ, USA.