TransInformação

ORIGINAL

Editor
Valéria dos Santos Gouveia Martins

Conflict of interest
The authors declare that there are no conflicts of interest.

Data Availability
The research data are available from the corresponding author upon reasonable request.

Received
11 Nov. 2024

Final version
10 Mar. 2025

Approved
17 Mar. 2025

Extracting the meaning of a word: an artificial intelligence approach

Extraindo o significado de uma palavra: uma abordagem de Inteligência Artificial

Aerty Pinto dos Santos[1], Eduardo Almeida Santos Oliveira[1], Juliana Pinheiro Campos Pirovani[2], Elias de Oliveira[1]

[1] Universidade Federal do Espírito Santo, Departamento de Arquivologia, Programa de Pós-Graduação em Informática. Vitória, ES, Brasil. Correspondence to: E. OLIVEIRA. E-mail: <elias@lcad.inf.ufes.br>.
[2] Universidade Federal do Espírito Santo, Curso de Ciência da Computação, Departamento de Computação. Vitória, ES, Brasil.

How to cite this article: Santos, A. P. *et al*. Extracting the meaning of a word: an artificial intelligence approach. *Transinformação*, v. 37, e2514829, 2025. https://doi.org/10.1590/2318-0889202537e2514829

Abstract

This article presents a strategy to extract the meaning of words in different contexts, using classification algorithms such as kNN, WiSARD, and 1NN, combined with a robust language model. The main objective is to investigate how the term "archive" is used in journalistic articles and how this usage reflects the value placed on the work of archivists. To achieve this, texts published in the newspaper "*A Tribuna*" between 2003 and 2017 were analyzed. The adopted method involves the automatic classification of sentences containing the term "archive," dividing them into eleven categories that represent different interpretations of the term. The research was conducted through a classification algorithm, trained to identify semantic patterns in the sentences. This is a textual data analysis extracted from a digital collection of a periodical, without the direct participation of human subjects. The results indicate that combining the language model with the neural network significantly improves classification performance, surpassing traditional methods in metrics such as precision and recall. Additionally, the analysis showed that the term "archive" is widely used in different contexts by journalists, revealing multiple meanings and highlighting the importance of archivists in the process of organizing and documenting records. The proposed approach shows potential for application in other domains, contributing to the automation of semantic inference and the classification of large volumes of textual data.

Keywords: Contextual meaning. Machine learning. Natural language processing. Semantic classification. Text analysis.

Resumo

Este artigo apresenta uma estratégia para extrair o significado de palavras em diferentes contextos, utilizando algoritmos classificadores, como kNN, WiSARD e 1NN, combinados com um modelo de linguagem robusto. O objetivo central é investigar como o termo "arquivo" é empregado em artigos jornalísticos e de que forma essa utilização reflete a valorização do trabalho de arquivistas. Para isso, foram analisados textos publicados no jornal "A Tribuna" entre os anos de 2003 e 2017. O método adotado envolve a classificação automática das frases que contêm o termo "arquivo", dividindo-as em onze categorias que representam diferentes interpretações desse termo. A pesquisa foi realizada por meio de um algoritmo

de classificação, que foi treinado para identificar padrões semânticos nas frases. Trata-se de uma análise de dados textuais extraídos de um acervo digital de um periódico, sem a participação direta de sujeitos humanos. Os resultados apontam que a combinação do modelo de linguagem com a rede neural melhora significativamente o desempenho da classificação, superando métodos tradicionais em métricas como precisão e recall. Além disso, a análise mostrou que o termo "arquivo" é amplamente utilizado em diferentes contextos pelos jornalistas, revelando múltiplos significados e ressaltando a relevância dos arquivistas no processo de organização e registro documental. A abordagem proposta demonstra potencial para ser aplicada em outros domínios, contribuindo para a automação da inferência semântica e a classificação de grandes volumes de dados textuais.

Palavras-chave: *Significado contextual. Aprendizado de máquina. Processamento de linguagem natural. Classificação semântica. Análise de texto.*

Introduction

Understanding the meaning of a particular word is an important part of many machine learning and natural language processing problems. A word can be used in various forms in a sentence and convey a different underlying message. For instance, the word archive, *arquivo* – in Portuguese, may express a place where someone lets in their documents: "John, all my reports are saved in the archive at my office". However, in the following sentence: "Could you please delete all the archives on my computer?" – the meaning of the word archive is rather different. The use of a word says a lot about the one who uses it: [...] the mouth speaks what the heart is full of (Luke 6:45).

Our investigation departs from the above understanding to unveil how other information professionals, e.g. journalists in the current work, frequently use the word archive when registering their work in newspapers. We would like to know whether these professionals value the work carried out by the archivists in organizing documents for the archival.

Therefore, our resource consists of documents obtained from a local newspaper, and the access to its daily digital issues is public. We then extract all the sentences from each newspaper's page where there are one or more occurrences of the word archive. Our proposed algorithm classifies each sentence into one of the 11 meaning classes suggested by Grigoleto, Aldabalde and Oliveira (2017) in their work.

In the cited study, the process was entirely manual. However, in the current proposal, we aim to extend that scope to increase the volume of documents, enhancing the statistical analysis of the current ongoing research. To address the increase in sentences, we propose a semi-automatic approach. That involves training an artificial intelligence algorithm with minimized effort, followed by its application to classify the meaning of words occurring in the remaining unseen sentences. We will call these sentences 'documents' from now on in this work.

The goal of this research is twofold. Firstly, we aim to develop methods for automatically classifying the meaning of a word within a sentence. In our case, we are interested in the meaning of the word archive (*arquivo*, in Portuguese) when signifying archival in Portuguese. Secondly, we want to assess the importance of the archivist job in the journalistic context. What is the impact of this kind of job for newspaper professionals? Furthermore, our research extends beyond linguistic exploration to practical implications. We aspire to map the archivist job market comprehensively, examining the evolving demands, skill sets, and professional landscapes within the field. By doing this, we aim to provide valuable insights into the current state and future trends of archiving professions, offering a nuanced understanding of the skills and competencies required for success in this dynamic and vital sector. Through this multifaceted approach, our research seeks to contribute to both theoretical and practical advancements in the realm of archiving and information management.

This article is structured as follows. In Section 2, we discuss some of recent work that will serve as a basis for comparison with our proposal. Our proposal to deal with the classification of the term archive is discussied in Section 3. The results are discussed in Section 4. In Section 5, the conclusions and future works are presented.

Literature Review

The use of the right concepts is key in information transferring and knowledge building. The author Azevedo Netto (2008) pointed out that this happens at the discursive level of each informational community. In our work, we are interested in extracting the meaning a newspaper journalist conveys when using the word archive in their discourses. One goal in this work is to gauge the importance these communities give to the main core of the archivist's activities.

Artificial Intelligence refers to the ability of machines and computers to perform tasks that normally require human intelligence. It involves the ability to learn, reason, solve problems, understand natural language and interact intelligently with the environment. Among several subcategories, there is natural language processing focused on the interaction between man and human language. Another branch of AI is machine-learning, which seeks to learn patterns to aid in prediction and classification. This classification is the objective of this study. There are several classification algorithms used in Machine Learning, of which we can mention those used in this work: kNN, WiSARD and 1NN.

The choice of kNN, WiSARD, and 1NN models for the semantic classification of the term "archive" was motivated by computational efficiency, interpretability, and adaptability to smaller datasets. WiSARD enables fast classification based on associative memory (De Gregorio *et al.*, 2022), while 1NN often outperforms more complex methods (Geler *et al.*, 2016). In contrast, SVM and Neural Networks require greater computational power and fine-tuning of hyperparameters (Pannakkong *et al.*, 2022), and Random Forest, although robust, may demand more feature engineering without a proportional gain (Gul *et al.*, 2018). Both kNN and WiSARD operate based on similarity; kNN classifies new elements by identifying the nearest neighbors in a feature space, while WiSARD evaluates patterns through the activation of specific memory positions (Kappaun *et al.*, 2016). Both provide an efficient and straightforward approach to the problem, balancing accuracy, performance, and ease of implementation.

Generative Pre-trained Transformer (GPT) technology was originally developed with the aim of advancing the field of Natural Language Processing, it is a language model architecture based on transformers, which are a class of machine learning models. Effectively and versatilely, these models are trained on large amounts of textual data to learn general language representations, allowing them to generate text, answer questions, and perform a variety of other natural language-related tasks, including classification (Brown *et al.*, 2020).

A recent study explored breast cancer prediction using the kNN algorithm in conjunction with Min-Max and Z-Score normalization methods. Conducted in the R language, the study compared the accuracy of these methods in distinguishing between malignant and benign cancers based on 569 biopsy samples and 32 variables from the Wisconsin Breast Cancer Diagnostic Dataset. The results showed that Min-Max normalization achieved an accuracy rate of 98% for k values of 5 and 21, slightly surpassing the Z-Score method, which reached a maximum accuracy of 97% for k values of 5 and 15 (Henderi, 2021). These findings highlight the efficiency of the kNN algorithm in achieving high classification accuracy rates, even with a relatively small dataset.

Another study (Reiss, 2023), also focusing on text classification, investigated the evaluation of the zero-shot capabilities' reliability of ChatGPT. These capabilities refer to the ability of a language model to perform a specific task without being explicitly trained for it. The methodology employed involved obtaining texts from German language websites, which were converted into plain text after analyzing the corresponding HyperText Markup Language (HTML), with the purpose of distinguishing between news and non-news texts from websites. To instruct ChatGPT in the classification task, ten distinct instructions were developed. The results indicated that the consistency in ChatGPT's classification output may not meet scientific standards of reliability. Krippendorff's Alpha was used as a metric in the article to assess whether the same input generates consistent outputs. Consistencies with a Krippendorff's Alpha above 0.8 are considered reliable; however, it was observed that slight variations in instructions or the repetition of identical inputs result in significant differences in outputs. Additionally, it is highlighted the need to consider effective instruction strategies for text annotation contexts, as the efficient approach for human annotators may not be ideal for ChatGPT, underscoring the importance of developing instruction strategies tailored to the characteristics of ChatGPT to optimize its performance in specific text classification tasks.

The Proposed Approach

The proposed approach, illustrated in Figure 1, involves acquiring only those pages containing the word *arquivo* from the "*A Tribuna*" newspaper archive. Subsequently, these pages will be subjected to an additional algorithm that will extract each sentence which contains the searched term "archive". These sentences will then be stored in a table, which will be used later as input for a GPT (Kublik; Saboo, 2023) language model with the goal of classifying the associated sentences.

In traditional supervised classification methods, a model receives input and produces an output from a fixed set of classes after training (Muhammad; Algehyne; Usman, 2020). However, Large Language Models, such as the GPT-3.5 Turbo used in this article, have introduced a new approach where an additional text specific to the task, called prompt, is included along with the model input. This prompt may contain questions about the current sample, input-output pair of examples, or task descriptions. The prompting approach eliminates the need for additional training in the GPT technology, though fine-tuning remains an option. This approach has been effective in various supervised classification applications (Chang *et al.*, 2024).

Therefore, part of our approach methodology is as follows: First we present to a GPT some illustrative examples to explain the desired intent of the classification process – some actual sentence with the word archive in it and the correspondent classification tag. We use the same classifications tags proposed by Grigoleto, Aldabalde and Oliveira (2017): c1: As an electronic document; c2: A set of documents; c3: As a paper document; c4: As an archival institution; c5: As the name of a product (e.g., a cultural product); c6: As a source (e.g., Source: "*A Tribuna*" Archive); c7: As the conclusion of a legal process (e.g., the case was archived); c8: As a storage and packaging fixture; c9: As a witness (e.g., the burning of files in a witness murder); c10: As a sector of a company; c11: As memory (e.g., an account of an unforgettable event).

Along with an explanation of the meaning of each class from c1 to c11 related to possible archival objects, the prompt will also include an example question and the expected response, such as:

User: opened the file (*arquivo* in Portuguese) using Microsoft PowerPoint2010, clicked on the Tools tab, and then clicked on the Convert option.

Assistant: c1.

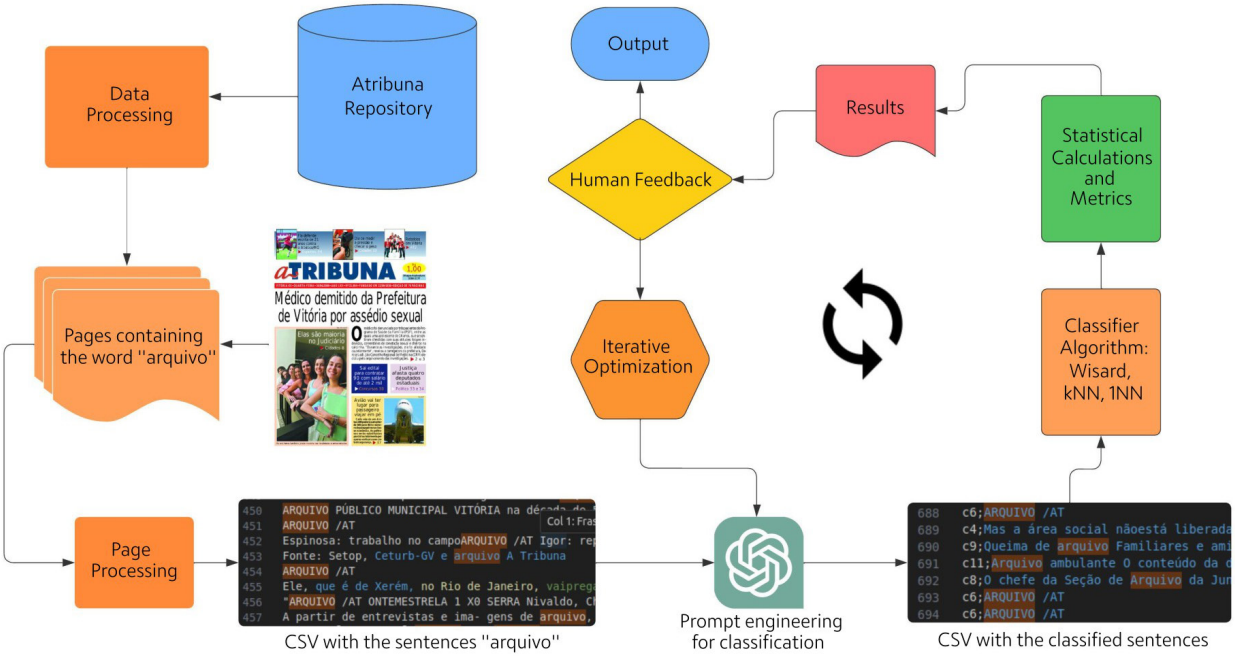

Figure 1 – Method of classifying Word arquivo.
Source: Elaborated by authors (2024).

Following this initial phase, a statistical metrics evaluation will be conducted to determine the effectiveness of the GPT model classification. This evaluation will involve comparing the results with the statistical metrics of the previously employed kNN, WiSARD and 1NN classification models. Metrics such as Precision, Recall, and F1-score will be considered for each model (Baeza-Yates; Ribeiro-Neto, 2011).

Based on the results of this evaluation, a human feedback process will be conducted to obtain qualitative insights into the performance of the different classification models. These insights will be crucial in guiding the decision-making regarding the continuation or termination of the development cycle.

Eventually, the user may present new inputs to the prompt to improve GPT performance. These inputs may include adjustments to model parameters, expansion of the training dataset, or refinement of text pre-processing strategies. The goal is to improve the development cycle incrementally. In doing so, we incorporate human feedback and statistical evaluations to optimize the GPT model's classification.

Experiments and Results

The text classification problem is still a challenging problem for the scientific communities. One of the problems is the cost of training an algorithm to learn to classify a large quantity of archive documents. These algorithms usually need a large set of previously labeled data (Li *et al.*, 2022). That means, therefore, a great effort of human annotations and cross-validations.

We propose an approach where the cost to train our algorithm is cheaper than the majority of the literature alternatives. In other words, our algorithm can quickly learn, and its quality

performance is statistically comparable to the baseline literature algorithms. To test our general approach, we described experiments and results yielded by our proposal when classifying the same dataset of 45.345 presented by the authors Oliveira and Branquinho Filho (2017).

The choice of evaluation metrics is essential for validating the performance of classification models in NLP Natural Language Processing. In this study, we use precision, recall, and F1-score, which are widely adopted in the literature for capturing different aspects of model effectiveness (Yacouby; Axman, 2020). Additionally, considering the influence of class imbalance on result analysis, we use macro and micro averages of these metrics to ensure a fairer and more representative evaluation (Riyanto *et al.*, 2023).

The precision metric is fundamental in evaluating similarity-based classifiers, as it measures the proportion of correct predictions relative to the total predictions made for a specific class. In the context of this study, precision is essential for assessing the model's accuracy in identifying the different meanings of the term "archive" in journalistic articles. Although similarity between samples is a relevant criterion in classification, proximity between examples does not guarantee correct predictions, as noisy data or complex patterns may lead to incorrect label assignments. Thus, precision not only evaluates the degree of similarity between samples but also verifies whether the model is learning useful patterns or merely capturing spurious relationships. Its formula is expressed as:

Recall, in turn, measures the model's ability to correctly identify relevant instances, reducing false negatives. In this study, this metric was used to evaluate the retrieval of sentences containing the term "archive", ensuring that different semantic interpretations were identified and correctly classified. This aspect is essential, as the research aims to map the recognition of archival work in the press. Recall is defined by the equation:

In multiclass classification tasks, the F1-score becomes particularly relevant, as it prevents the overvaluation of majority classes and ensures a balanced evaluation across all categories (Riyanto *et al.*, 2023). In this study, this metric is essential for assessing the model's performance in distinguishing the eleven categories assigned to the term "archive", ensuring that no meaning is overlooked due to unequal data distributions. To balance precision and recall, we use the F1-score, which represents the harmonic mean of these metrics and is widely recommended for scenarios with class imbalance:

In addition to these metrics, accuracy was considered to assess the overall proportion of correct classifications. Defined as the ratio between the number of correctly classified instances and the total number of instances, its formula is expressed as:

Although useful in balanced class contexts, accuracy can be misleading in imbalanced scenarios, as a model may achieve high scores simply by classifying all instances as belonging to the majority class. Thus, metrics such as precision, recall, and F1-score provide a more detailed and reliable assessment of model performance.

According to Buttcher, Clarke and Cormack (2016), evaluating the effectiveness of a classification method depends on the perceived relevance in human analyses. Therefore, the combined use of these metrics allows for a more comprehensive understanding of the model's ability to correctly differentiate classes and minimize classification errors.

Table 1 shows our results on classifying a large dataset of 45.345 documents. We improved the results in all the metrics adopted by the authors (Oliveira; Branquinho Filho, 2017) by 18% on the F1 metric. Although the importantance of the latest metric result is an average of two others, we are

mainly interested in the precision yielded by the WiSARD approach of 95%, against 77% achieved by that proposed by the previously mentioned authors.

Table 1 – Comparison between Micro-averaged and Macro-averaged F1 Scores.

Algorithm	Accuracy	Precision	Recall	F2
Micro-averaged scores				
1NN	0.821	0.895	0.821	0.842
kNN	0.845	0.930	0.845	0.879
WiSARD	0.943	0.943	0.943	0.943
Macro-averaged scores				
1NN	0.821	0.883	0.812	0.831
kNN	0.821	0.883	0.812	0.831
WiSARD	0.943	0.958	0.953	0.950

Source: Elaborated by authors (2024).

Given these promising results, we now look into the classification of a single word within a sentence chosen for this research. An example of a sentence is as follows:

And therein the archive, they had direct contact with the records (Original sentence: *E ali no arquivo eles tinham contato direto com os registros*)[3].

In this case, the meaning refers to a public organization named *Arquivo Público Estadual*. Whereas, in the following sentence the meaning is rather different:

The file will be moved to a floppy disk. (Original sentence: *O arquivo será movido para um disquete*)[4].

Note that in English the word is already another, but in Portuguese is the same. In English, the word make it clear that we are talking about something that can exists within a computer equipment.

Using the suggested methodology, the development of the bar graph highlighted in Figure 2 was carried out. This graph represents the supervised classifications from c1 to c11 assigned to sentences extracted from the pages of the "*A Tribuna*" newspaper.

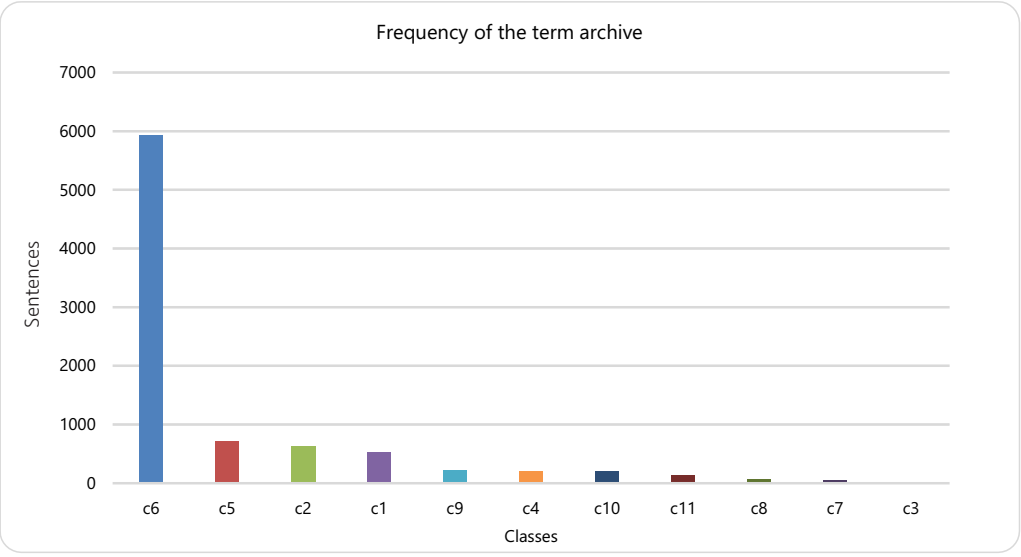

Figure 2 – Frequency of use of the term arquivo.
Source: Elaborated by authors (2024).

[3] http://www.qsabe.com.br/dataSets/pagina12.pdf.
[4] http://www.qsabe.com.br/dataSets/pagina18.pdf.

In Figure 2, a significant prominence of categories c6, c5, c2, and c1 is observed, while the term *arquivo*, represented by class c4 (archival institution), appears less frequently. This result reinforces the idea that the use of the term *arquivo* is largely influenced by common sense, reflecting the limited democratization of these institutions. Additionally, it highlights the low presence of custodial institutions in the media, which affects the perception of public and private archives in the regional journalistic context.

The initial analysis of Figure 3 revealed variability in classification results, demonstrating that the same entry can be categorized in different ways. Ideally, a sentence containing the term "ARQUIVO/AT" should consistently be classified as c6 (source). However, it was observed that ChatGPT tends to analyze the surrounding context of the term to determine its classification. Thus, even though the mere presence of the term characterizes the sentence as c6, the inclusion of words related to information technology may lead to its classification as c1 (electronic document).

Furthermore, the inherent randomness of the ChatGPT model, intensified by its black-box nature, can compromise the reliability of the final classification. Despite efforts to optimize and control variables through validation, discrepancies persist, influenced by factors such as temperature, which regulates the degree of randomness in responses, and the formulation of the input prompt itself.

When submitting the sentences, as illustrated in Figure 3, to two distinct classifications using different prompts, comparative results with the supervised data were obtained. In Classification 1, class c6 (source) showed a significant reduction compared to the classification of supervised data. A considerable number of errors associated with this class were recorded, attributed to the non-deterministic nature of ChatGPT, which takes the context around the term arquivo into account, posing challenges for classification. On the other hand, classes c3 (paper document), c5 (cultural product), and c11 (memory) showed a substantial increase, with values that in the supervised data originally belonged to class c6. In this approach, the same classification tags proposed by Grigoleto, Aldabalde and Oliveira (2017) were used in the prompt without applying any preprocessing to the set of sentences.

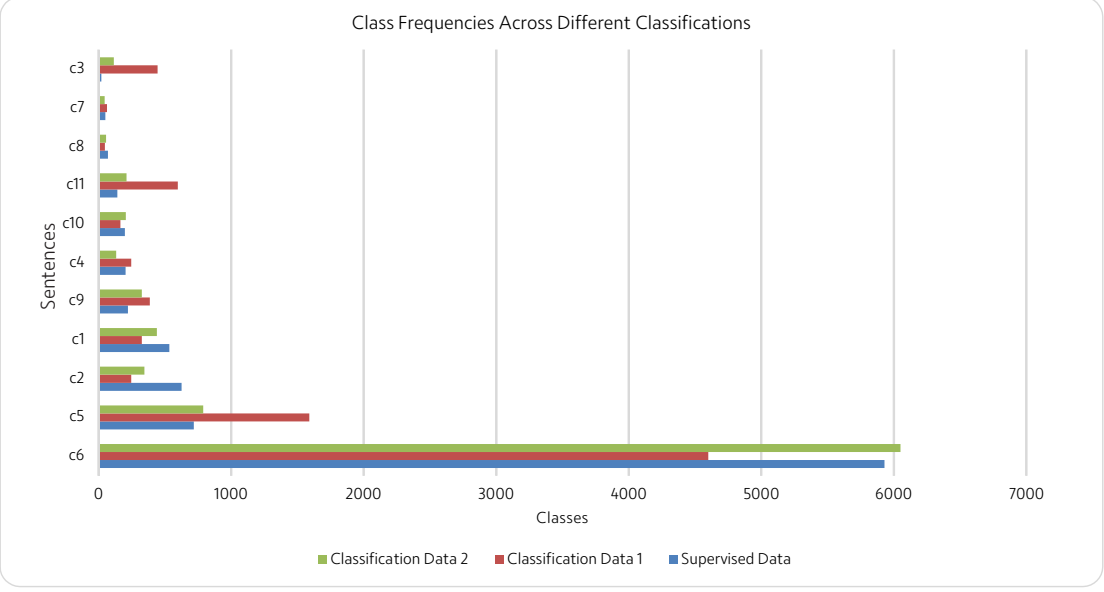

Figure 3 – Frequency of classes in different classifications.
Source: Elaborated by authors (2024).

Conversely, in Classification 2, shown in Figure 3, preprocessing was applied to the sentences containing the term "ARQUIVO/AT", keeping only this term. Additionally, more examples of classification scenarios were added to the prompt. These interventions resulted in a graph more aligned with the supervised data, indicating an improvement in classification accuracy. This process helped minimize the previously observed discrepancies in classes c6, c3, c5, and c11, emphasizing the importance of proper data preprocessing and precise prompt formulation to obtain more reliable results that closely align with supervised data.

By employing the sentences classified via GPT in the WiSARD classifier, the results highlighted in Figure 4 were observed. In the macro F1 metric of this classifier, a score of 64% was achieved with supervised data. However, considering the data from the first classification, a score of 66% was obtained, while the second classification data resulted in a score of 67%. The analysis of the metrics and the graph in Figure 2 suggests that the better these results, the closer the data classification will be to that achieved through a prompt with supervised classification examples.

These results suggest that using sentences classified by GPT in the WiSARD classifier can provide valuable insights compared to using supervised data. Analyzing the metrics obtained by WiSARD at each classification stage may reveal significant opportunities for improvement. Careful prompt adjustment and the selective choice of examples can not only bring the results closer to those of supervised models but may even surpass them, considering the various possible interpretations of the word archive.

The Analysis of Variance (ANOVA) test (Morettin, 2010) is a widely used statistical technique for comparing the means of different groups and determining whether there are statistically significant differences among them. In the context of this study, ANOVA was applied to evaluate the performance of the 1NN, WiSARD, and kNN classifiers based on the metrics Accuracy, Precision, Recall, and F1-score, considering both micro-average and macro-average. The micro-average weights the results according to the number of instances in each class, while the macro-average calculates the simple mean of the metrics per class, treating all classes equally (Nurrahma; Yusuf, 2020).

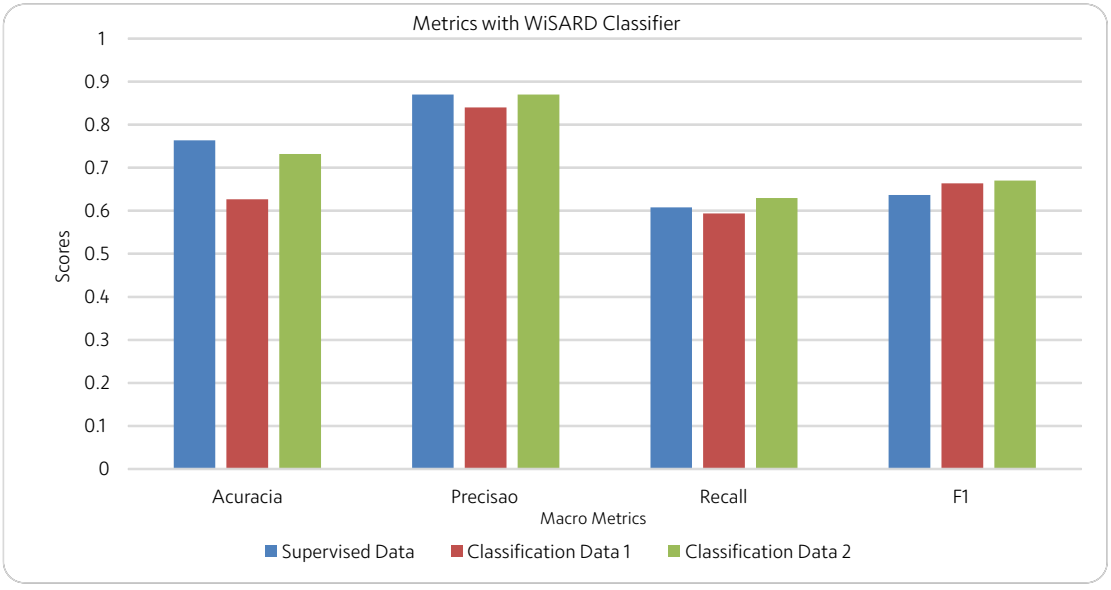

Figure 4 – Macro Metric of the WiSARD Classifier.
Source: Elaborated by authors (2024).

The null hypothesis of ANOVA states that there is no significant difference between the means of the compared groups. For this study, a significance level of $p < 0.05$ was considered. If the *p*-value is below this threshold, the null hypothesis is rejected, indicating that at least one group differs significantly from the others. The ANOVA results indicated no statistically significant differences among the classifiers for Accuracy ($p = 0.5903$), Recall ($p = 0.2133$), and F1-score ($p = 0.1182$), suggesting similar performance across these metrics. However, for Precision ($p = 0.0009$), the null hypothesis was rejected, revealing significant differences among the classifiers, particularly in the macro-average.

Given this result, a Tukey post hoc test (Nogueira *et al.*, 2020) was conducted to identify which classifier pairs exhibited statistically significant differences. The analysis revealed that the 1NN classifier achieved significantly higher Precision compared to WiSARD ($p = 0.0007$) and kNN ($p = 0.0072$), while no significant difference was observed between WiSARD and kNN ($p = 0.083$). These findings suggest that, particularly in the macro-average, 1NN demonstrates more consistent and superior performance in terms of Precision, making it the most suitable choice when this metric is a priority in classifier evaluation.

Conclusion

This study proposed an innovative strategy to automatically infer the meaning of the word "archive" in context by leveraging a large language model for classification and using classifier algorithms to validate the results with metrics. The approach combines the strengths of both: the language model provides a deep contextual understanding of the text, while the classifier algorithms assess the quality of the classification.

In the experiments, we aimed to determine the meaning of the word "archive" in various newspaper articles. By analyzing its usage, we observed that it can refer both to an archival institution and a physical location. Our goal was to understand the relevance of the archive as an institution and its role as social memory in the work of journalists. Surprisingly, the meaning initially considered the most relevant ranked only sixth in frequency (c4), while the category (c6), related to internal archives, appeared in first place, highlighting its importance for these professionals.

The use of sentences generated by language models in classifiers proved to be a promising alternative to supervised methods. While traditional approaches remain a solid benchmark, data generated by the language model, when applied to classifiers like WiSARD, provided valuable insights and improved classification metrics. With appropriate adjustments and a careful selection of examples, this approach can match or even surpass supervised classifications, particularly in contexts where a single term carries multiple interpretations.

Moreover, combining the language model with classifiers significantly reduced the human effort required to train the algorithm, making this approach ideal for handling large datasets. For future research, we plan to extend this strategy beyond journalistic articles, broadening its applications in automated text classification.

Despite the method's effectiveness, it is essential to consider the limitations of the corpus used. The analysis was based exclusively on texts from "*A Tribuna*" newspaper (2003-2017), reflecting its editorial style, regional context, and specific timeframe. Since different media outlets may use and interpret the term "archive" in distinct ways, the generalizability of the findings may be affected. To validate the results more comprehensively, future studies should include texts from various publications, covering different regions, historical periods, and editorial perspectives.

Finally, while our focus was on the term *arquivo* in journalistic articles, its meaning varies across text genres. In news reports, it may relate to information sources (c6) or archival institutions (c4). In academic papers, to document collections (c2) or electronic records (c1). In social media, to memory (c11) or digital documents (c1). In technical documents, to storage systems (c8). These variations suggest that future research could apply the proposed methodology to different text genres, further exploring the polysemy of the term in contemporary language.

References

Azevedo Netto, C. X. A abordagem do conceito como uma estrutura semiótica. *Transinformação*, v. 20, n. 1, 2008.

Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval*. 2nd. ed. New York: Addison-Wesley, 2011.

Brown, T. *et al*. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, v. 33, p. 1877-1901, 2020. Doi: https://doi.org/10.48550/arXiv.2005.14165.

Buttcher, S.; Clarke, C. L. A.; Cormack, G. V. *Information retrieval*: Implementing and evaluating search engines. Cambridge: MIT Press, 2016.

Chang, K.-W. *et al*. SpeechPrompt: Prompting Speech Language Models for Speech Processing Tasks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 32, p. 3730-3744, 2024. Doi: https://doi.org/10.1109/TASLP.2024.3436618.

De Gregorio, M. *et al*. Classification of preclinical markers in Alzheimer's disease via WiSARD classifier. *In*: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 30., 2022, Bruges. *Proceedings* [...]. Bruges: ENNS, 2022. p. 43-48. Doi: https://doi.org/10.14428/esann/2022.ES2022-63.

Geler, Z. *et al*. Comparison of different weighting schemes for the kNN classifier on time-series data. *Knowledge and Information Systems*, v. 48, p. 331-378, 2016. Doi: https://doi.org/10.1007/s10115-015-0881-0.

Grigoleto, M. C.; Aldabalde, T. V.; Oliveira, E. Discutindo a questão da polissemia do termo arquivo na imprensa: um estudo a partir da Teoria do Continuum. *In*: Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB), 17., 2017, Rio de Janeiro. *Anais* [...]. Rio de Janeiro: ANCIB, 2017.

Gul, A. *et al*. Ensemble of a subset of kNN classifiers. *Advances in Data Analysis and Classification*, v. 12, n. 4, p. 827-840, 2018. Doi: https://doi.org/10.1007/s11634-015-0227-5.

Henderi, H. Comparison of min-max normalization and z-score normalization in the k-nearest neighbor (knn) algorithm to test the accuracy of types of breast cancer. *IJIIS*: *International Journal of Informatics and Information Systems*, v. 4, p. 13-20, 2021. Doi: https://doi.org/10.47738/ijiis.v4i1.73.

Kappaun, A. *et al*. Evaluating Binary Encoding Techniques for WiSARD. *In*: Brazilian Conference on Intelligent Systems (BRACIS), 5., Recife, 2016. *Proceedings* [...]. Recife: SBC, 2016. p. 103-108. Doi: https://doi.org/10.1109/BRACIS.2016.029.

Kublik, S.; Saboo, S. GPT-3: The ultimate guide to building NLP products with OpenAI API. [*S. l.*]: Packt Publishing, 2023.

Li, Q. *et al*. A Survey on text classification: from traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology* (*TIST*), v. 13, n. 2, p. 1-41, 2022. Doi: https://doi.org/10.1145/1122445.1122456

Morettin, P. A.; Bussab, W. O. *Estatística básica*. 6. ed. São Paulo: Saraiva, 2010.

Muhammad, L. J.; Algehyne, E. A.; Usman, S. S. Predictive supervised machine learning models for diabetes mellitus. *SN Computer Science*, v. 1, n. 5, 2020. Doi: https://doi.org/10.1007/s42979-020-00250-8.

Nogueira, C. J. *et al*. Amplitude de movimento de militares submetidos a 12 semanas de alongamento com diferentes intensidades. *Educación Física y Ciencia*, v. 22, p. 3, e135, 2020. Doi: https://doi.org/10.24215/23142561e135.

Nurrahma, R.; Yusuf, R. Comparando diferentes precisões de aprendizado de máquina supervisionado na análise de dados da COVID-19 usando o teste ANOVA. *In*: *International Conference on Interactive Digital Media (ICIDM)*, 6., 2020, Bandung. *Proceedings* [...]. Bandung: UTM, 2020. p. 1-6. Doi: https://doi.org/10.1109/ICIDM51048.2020.9339676.

Oliveira, E.; Branquinho Filho, D. Automatic classification of journalistic documents on the Internet. *Transinformação*, v. 29, n. 3, 2017. Doi: https://doi.org/10.1590/2318-08892017000300003.

Pannakkong, W. *et al.* Hyperparameter tuning of machine learning algorithms using response surface methodology: a case study of ANN, SVM, and DBN. *Mathematical Problems in Engineering*, v. 2022, p. 1-17, 2022. . Doi: https://doi.org/10.1155/2022/8513719.

Reiss, M. V. Testing the reliability of ChatGPT for text annotation and classification: a cautionary remark. *arXiv*, 2023. Doi: https://doi.org/10.48550/arXiv.2304.11085.

Riyanto, S. *et al.* Comparative analysis using various performance metrics in imbalanced data for multi-class text classification. *International Journal of Advanced Computer Science and Applications*, v. 14, n. 6, 2023. Doi: http://dx.doi.org/10.14569/IJACSA.2023.01406116.

Yacouby, R.; Axman, D. Probabilistic extension of precision, recall, and F1 score for more thorough evaluation of classification models. *In*: *Workshop on Evaluation and Comparison of NLP Systems*, 1., 2020. *Proceedings* [...]. [*S. l.*]: Association for Computational Linguistics, 2020. p. 79-91. Doi: https://doi.org/10.18653/v1/2020.eval4nlp-1.9.

Contributors

Conceptualization: E. OLIVEIRA. Data curation: A. P. SANTOS. Investigation: E. A. OLIVEIRA. Project administration: E. OLIVEIRA. Software: A. P. SANTOS and E. A. OLIVEIRA. Supervision: E. OLIVEIRA. Validation. J. P. C. PIROVANI. Writing – original draft: E. OLIVEIRA. Writing – review and editing: A. P. SANTOS, J. P. C. PIROVANI and E. OLIVEIRA.