Leveraging Large Language Models for Author Name Disambiguation in Portuguese Contexts

Samuel G. dos Santos[1], **Vitória M. Diniz**[1]
Bartolomeu S. Gusella[1], **Natan de S. Rodrigues**[1]

[1]Academic Institute of Technological Sciences,
State University of Goiás, GO – Brazil

samuel.santos@aluno.ueg.br, natan.rodrigues@ueg.br

Abstract. *Author Name Disambiguation (AND) is a fundamental task in digital libraries and repositories, especially in Portuguese contexts where metadata often lacks persistent identifiers and shows frequent inconsistencies. This study presents an unsupervised approach that integrates Large Language Models (LLMs) for semantic normalization, MiniLM embeddings for similarity modeling, and automatic clustering followed by a post-merging heuristic. Experiments on the BDBComp dataset show competitive cluster cohesion (K = 0.907) compared to baselines, while the pairwise F1 score (pF1 = 0.448) highlights the difficulty posed by highly ambiguous surnames. Future work will refine LLM summaries and clustering thresholds to improve accuracy while preserving cluster consistency.*

1. Introduction

Digital libraries and institutional repositories depend on the correct identification of authors to support bibliometric analyses, collaboration networks, and research evaluation. The Author Name Disambiguation (AND) task addresses this challenge by defining which works belong to each researcher. Ambiguity arises when different researchers share the same name or when one researcher publishes with multiple variants. Incomplete metadata and the lack of persistent identifiers intensify this problem [Ferreira et al. 2012].

Portuguese repositories face these issues in a critical way. A study of eight polytechnic repositories showed that more than 90% of records lacked identifiers such as ORCID[1] or *Ciência ID*[2] and also presented frequent errors in name forms and truncation [Rodrigues and Rodrigues 2024]. These problems compromise retrieval and aggregation and reinforce the demand for automated solutions.

Researchers developed heuristic, supervised, and hybrid approaches to address the AND task. The literature review in [Rodrigues et al. 2024] analyzed 211 studies from 2003 to 2022 and confirmed the predominance of clustering strategies while noting the absence of semantic enrichment and Large Language Models (LLMs). Recent studies tested LLMs in AND pipelines [Zhang et al. 2024, Yan and AsirAsir 2024, Zhao and Chen 2025], but these works rely on supervision, external retrieval, and experiments focused on English datasets.

[1] `https://orcid.org/`

[2] `https://www.ciencia-id.pt`

This study presents an unsupervised approach for AND in Portuguese bibliographic data. The proposal uses LLMs for semantic normalization, generates embeddings for similarity modeling, and applies clustering with a lightweight post-merging heuristic.

2. Related Work

The BDBComp dataset, introduced in [Cota et al. 2010], established a benchmark in Portuguese for evaluating AND methods. Early studies used unsupervised clustering with coauthor and venue information. Later works proposed hybrid and semi-supervised approaches, including associative classifiers [Ferreira et al. 2012] and the self-training strategy SAND [Ferreira et al. 2014]. These studies validated BDBComp as a reference but also showed limitations when metadata was sparse. Other research investigated AND in Portuguese contexts beyond BDBComp. The work in [Rodrigues et al. 2021] combined string similarity and network clustering, using DBLP data and examples from Brazilian graduate programs. This approach highlighted the need to adapt techniques to challenges typical of the Portuguese language, such as compound surnames and connectors.

LLMs emerged as a recent direction for AND. Iterative refinement with ChatGLM [Zhang et al. 2024], the integration of instruction-tuned models with LightGBM [Yan and AsirAsir 2024], and multilingual retrieval strategies [Zhao and Chen 2025] improved performance in English benchmarks but required supervision. In contrast, the present study applies LLMs as semantic cleaners in an unsupervised approach for Portuguese repositories.

3. Proposed Approach

The proposed approach addresses the AND task by grouping publication records into clusters representing unique authors. Figure 1 illustrates the workflow divided into four stages:

1. *Semantic Normalization with LLM:* metadata fields such as title, authors, and venue are processed by DeepSeek-R1:14B [DeepSeek-AI 2025] to generate short biographical summaries (*shortbios*). These summaries reduce noise and standardize the representation of each record.
2. *Sentence Embeddings:* the shortbios are encoded with the Sentence-Transformers model *paraphrase-MiniLM-L6-v2* [Wang et al. 2021], producing 384-dimensional vectors that capture semantic similarity between records.
3. *Clustering with Automatic Selection:* embeddings are grouped into author-level clusters using hierarchical agglomerative clustering (HAC), Leiden, and HDBSCAN. All methods are applied to each block, and the one with the best result is selected according to internal evaluation metrics.
4. *Post-merging Heuristic:* very small clusters are merged with their closest neighbor when sufficiently similar. This step reduces fragmentation and recovers valid pairs.

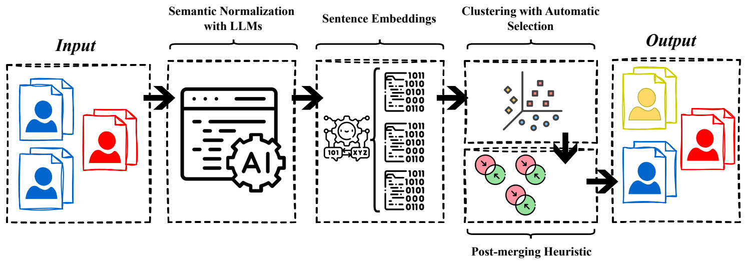
Semantic Normalization with LLMs
Sentence Embeddings
Clustering with Automatic Selection
Input
Output
AI
Post-merging Heuristic

Figure 1. Workflow of the proposed approach with four stages.

Experiments use the BDBComp dataset. Each record includes attributes such as *title*, *authors*, *venue*, and *year*. The collection is organized into ambiguous groups defined by common surnames, each containing multiple distinct authors. Table 1 summarizes the main groups, their sizes, and the number of real authors. The original dataset[3] lacked normalization. A normalized version and the complete implementation of the proposed approach are publicly available[4].

Table 1. Groups in BDBComp with number of records and distinct authors.

Group	# Records	# Authors	Group	# Records	# Authors
A. Oliveira	52	16	A. Silva	64	32
F. Silva	24	11	J. Oliveira	39	15
J. Silva	63	22	J. Souza	35	11
L. Silva	31	10	M. Silva	24	11
R. Santos	20	9	R. Silva	28	20

4. Preliminary Results

The proposed approach was evaluated on BDBComp dataset using pairwise F1 (pF1), K, and B^3-F1, three metrics commonly applied in AND [Ferreira et al. 2020]. The system used B^3-F1 to select the best clustering method for each author block, since this metric captures both precision and recall at the entity level.

For comparison with baselines, only pF1 and K are reported, as earlier studies on BDBComp did not include B^3-F1. With the automatic configuration, the approach tested all clustering methods described in Section 3 and selected the best result for each block. On average, it achieved pF1 = 0.448 and K = 0.907. HAC was chosen in most blocks, with the cut threshold set to the 0.60 quantile of 1-NN cosine distances, while Leiden was selected once with $k = 10$.

Table 2 summarizes the results. The proposed approach reaches K values close to baseline methods, although pF1 remains lower, particularly for highly ambiguous surnames such as *Silva* and *Souza*. This behavior reflects over-splitting and merging errors in challenging blocks. All experiments ran on a MacBook Air M1 with 16GB RAM and 256GB SSD under macOS Sequoia 15.3.2.

These results show that the approach already produces clusters with cohesion and coverage comparable to state-of-the-art baselines, while pairwise accuracy remains lower.

[3] `https://github.com/kashak79/DISC_AND_Data_Set/blob/master/bdbcomp.tar`

[4] `https://github.com/natansr/llmand`

Table 2. Average results on BDBComp: comparison restricted to pF1 and K, the metrics reported in baselines.

Method	pF1	K
Ours	0.448	0.907
HHC [Cota et al. 2010]	0.650	0.930
K-Way [Cota et al. 2010]	0.710	0.930
SAND-1 [Ferreira et al. 2014]	0.680	0.897
SAND-2 [Ferreira et al. 2014]	0.752	0.924

Future work will refine the semantic normalization performed by the LLM and explore adaptive clustering thresholds and post-merging strategies to improve recall without reducing cluster purity.

References

Cota, R. G., Ferreira, A. A., Nascimento, C., Gonçalves, M. A., and Laender, A. H. F. (2010). An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. *Journal of the American Society for Information Science and Technology*, 61(9):1853–1870.

DeepSeek-AI (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Ferreira, A. A., Gonçalves, M. A., and Laender, A. H. F. (2012). A brief survey of automatic methods for author name disambiguation. *Acm Sigmod Record*, 41(2):15–26.

Ferreira, A. A., Gonçalves, M. A., and Laender, A. H. F. (2020). *Automatic Disambiguation of Author Names in Bibliographic Repositories*. Morgan & Claypool Publishers.

Ferreira, A. A., Veloso, A., Gonçalves, M. A., and Laender, A. H. F. (2014). Self-training author name disambiguation for information scarce scenarios. *Journal of the Association for Information Science and Technology (JASIST)*, 65(6):1257–1278.

Rodrigues, M. E. P. and Rodrigues, A. M. (2024). Desambiguação de nomes de autores: um desafio para os repositórios / Disambiguation of authors' names: A challenge for repositories. *Ciência da Informação*, 53(3). 15ª Conferência Lusófona de Ciência Aberta (ConfOA), Modalidade: Pecha Kucha. IPCB & CERNAS-IPCB; ESA/IPCB & CERNAS-IPCB.

Rodrigues, N. S., Costa, A. R., Lemos, L. C., and Ralha, C. G. (2021). Multi-strategic approach for author name disambiguation in bibliography repositories. In Lossio-Ventura, J., Valverde-Rebaza, J., Díaz, E., and Alatrista-Salas, H., editors, *Information Management and Big Data. SIMBig 2020. Communications in Computer and Information Science, vol. 1410*. Springer, Cham.

Rodrigues, N. S., Mariano, A. M., and Ralha, C. G. (2024). Author name disambiguation literature review with consolidated meta-analytic approach. *International Journal on Digital Libraries*, pages 765–785.

Wang, W., Bao, H., Huang, S., Dong, L., and Wei, F. (2021). Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers.

Yan, Q. and AsirAsir (2024). Synergizing large language models and tree-based algorithms for author name disambiguation. In *Submitted to KDD 2024 OAG-Challenge Cup*.

Zhang, X., Zhou, Y., Chen, H., Bao, M., and Yan, P. (2024). Enhanced name disambiguation via iterative self-refining with LLMs. In *Submitted to KDD 2024 OAG-Challenge Cup*.

Zhao, R. and Chen, Y. (2025). Scholar name disambiguation with search-enhanced llm across language.