

DATA RESEARCH

meetup by MagIC

TaxonomAI:

Leveraging Large Language Models for Automated Taxonomy Construction

Farina Pontejos

NOVA Information Management School
Universidade NOVA de Lisboa
Lisbon, Portugal

INTRODUCTION

Automated Taxonomy Construction techniques typically made use of clustering algorithms such as K-means and Hierarchical Clustering.

With the rising popularity of Large Language Models (LLMs) however, the use of prompting as a zero-shot approach is emerging as a new field of research.

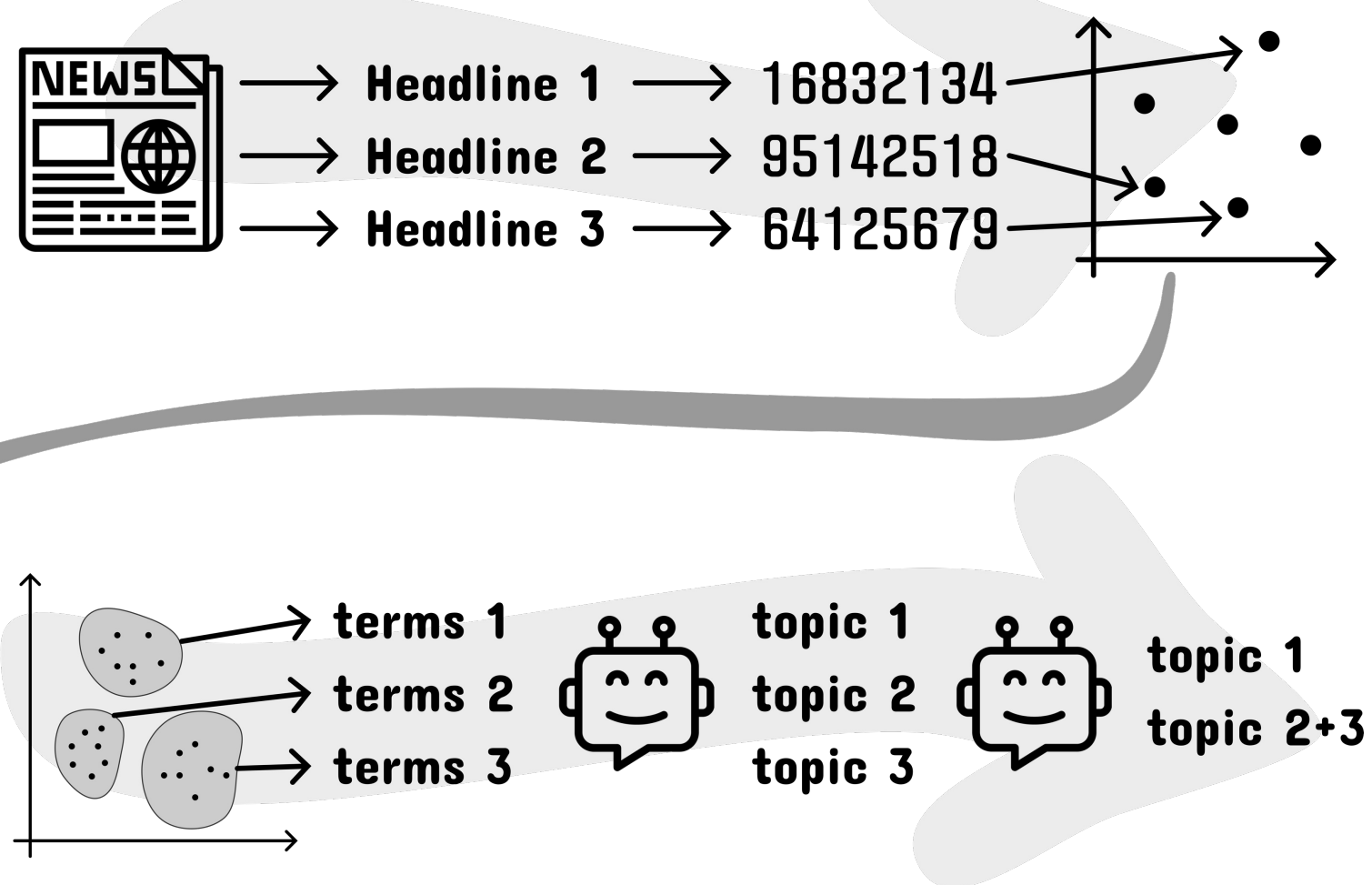
Approaches	Clustering	LLM Prompting
Challenges	Frequency-based	Training cost
	Loss of context	Hallucination
Advantages	Domain flexibility	Human-readable
Proposal	Frequency-based terms from document clusters	+ Translate terms into human-readable labels

METHODS AND MATERIALS

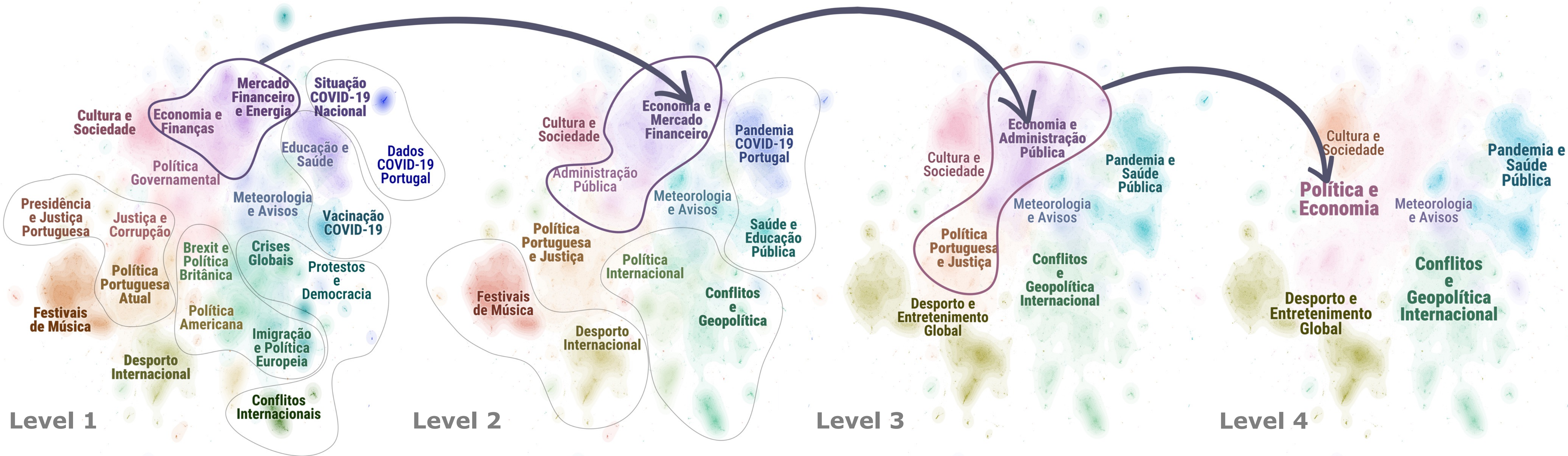
We extracted headlines from the front page snapshots of the news website [expresso.pt](https://www.expresso.pt) using the arquivo.pt API.

The text was vectorized using Sentence-BERT. Using Gaussian Mixtures, we derived 20 clusters, and with BERTopic derived keywords for each.

ChatGPT was then prompted to suggest better topic labels given these keywords, as well as suggest which topics can be merged.



RESULTS & DISCUSSION



Going from the lowest level (Level 1), we can see **Economia e Finanças** and **Mercado Financeiro e Energia** highlighted.

These topics are merged into **Economia e Mercado Financeiro** in Level 2, which in turn is merged with **Administração Pública** into **Economia e Administração Pública** in Level 3. Finally, this is merged with **Política Portuguesa e Justiça** into the top-most level as **Política e Economia**.

The LLM suggests which topics to merge only considering the keywords from the previous clusters, and occasionally suggested clusters that were distant in the embedding space.

We developed an interactive application to facilitate users' exploration of the corpus, accessible through the QR code. By visualizing the documents in two dimensions, the user can validate whether the topics being suggested for merging are indeed similar.



REFERENCES

- Zhang, C., Tao, F., Chen, X., Shen, J., Jiang, M., Sadler, B., Vanni, M., & Han, J. (2018). TaxoGen: Unsupervised Topic Taxonomy Construction by Adaptive Term Embedding and Clustering. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. doi:10.1145/3219819.3220064
- Wan, M., Safavi, T., Jauhar, S. K., Kim, Y., Counts, S., Neville, J., Suri, S., Shah, C., White, R. W., Yang, L., Andersen, R., Buscher, G., Joshi, D., & Rangan, N. (2024). TnT-LLM: Text Mining at Scale with Large Language Models (Version 1). *arXiv*. doi:10.48550/ARXIV.2403.12173

CONCLUSION

Using LLMs in ATC is a promising area of research, however more work is needed to determine how to quantitatively evaluate the resulting taxonomy.

Future research should focus on enhancing the integration of expert knowledge and exploring hybrid approaches that combine the strengths of both automated and manual methods to create more robust and dynamic taxonomies.

ACKNOWLEDGEMENTS

This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), under the project - UIDB/04152/2020 (DOI: 10.54499/UIDB/04152/2020) - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS).