Automated Phenotype-Based Clustering of Clinical Reports Using Large Language Models

Martina Saletta[1(✉)], Andrea Bombarda[1], Matteo Bellini[2], Lucrezia Goisis[2], Paolo Cazzaniga[1], Maria Iascone[2], and Domenico Fabio Savo[1]

[1] University of Bergamo, Bergamo, Italy
{martina.saletta,andrea.bombarda,paolo.cazzaniga,domenicofabio.savo}@unibg.it
[2] Laboratory of Medical Genetics, ASST Papa Giovanni XXIII, Bergamo, Italy
{m.bellini,miascone}@asst-pg23.it, lucreziagoisis@hotmail.it

Abstract. Large Language Models (LLMs) have shown significant potential in natural language processing tasks, including various applications in clinical and biomedical domains. This study explores the use of LLMs for analyzing a real dataset from Italian clinical reports and proposes a pipeline for automatically clustering these reports based on the described symptoms. The pipeline incorporates two approaches: (1) direct analysis of textual descriptions in the clinical reports, and (2) standardized processing through the automatic extraction of Human Phenotype Ontology terms using LLM-based methods. The obtained clusters will serve as the foundation for further predictive analyses, such as estimating the likelihood of a patient carrying specific genetic mutations. Our investigation compares the performance of direct text analysis against phenotype-standardized descriptions, highlighting the strengths and limitations of each approach.

Keywords: Large Language Models · Phenotype Clustering · Human Phenotype Ontology · Clinical Reports

1 Introduction

Clinical reports written in natural language contain a wealth of valuable information that can assist medical professionals in diagnosing conditions, identifying patterns, and making informed decisions. However, the unstructured nature of these documents poses challenges for their automated analysis. Large Language Models (LLMs) have recently emerged as powerful tools for processing and extracting meaningful insights from unstructured textual data, offering new possibilities for applications in clinical and biomedical domains [2,3].

In this work, we aim to analyze genetic testing reports produced in a clinical laboratory, by leveraging the k-means clustering algorithm to group reports

R. Bellazzi et al. (Eds.): AIME 2025, LNAI 15735, pp. 345–350, 2025.
https://doi.org/10.1007/978-3-031-95841-0_64

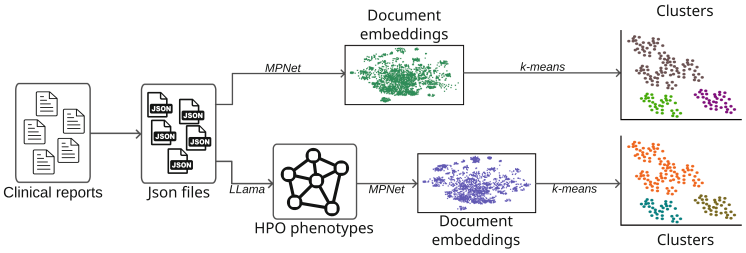

Fig. 1. Overview of the two approaches tested for document clustering.

based on their content. Specifically, k-means is applied to the embedding generated by the MPNet model [12]. The ultimate goal is to provide actionable insights to geneticists by identifying patterns and trends within the data.

To this end, we devise two approaches, which are outlined in Fig. 1: in the first one we work directly on the text as it appears in the document, by asking the LLM to embed the sentence describing the symptoms; the second one comprises an additional step aimed at recognizing and standardizing the described phenotypes according to the Human Phenotype Ontology (HPO) standard [5], and the LLM is asked to embed such phenotypes.

By comparing these two approaches, we assess their ability to produce clinically meaningful clusters and identify potential strengths and limitations. Our preliminary experiments, implemented in Python and discussed in Sect. 2, were conducted on a dataset of real medical reports, and the results show both approaches effectively cluster reports in clinically meaningful ways, despite some issues that will be discussed in Sect. 3.

To facilitate reproducibility and further research, the source code, data, and detailed results are available online [11].

2 Proposed Approach

In this work, we employed document clustering to analyze a dataset of medical reports comprising approximately 8,000 fully anonymized clinical reports documenting genetic tests performed at the Laboratory of Medical Genetics of ASST Papa Giovanni XXIII hospital in Bergamo, Italy. Each report is written in Italian and stored as a `.docx` file. They contain key information about the genetic test, including manifested symptoms, and test results. The semi-structured format of these files enabled us to preprocess the corpus into a collection of JSON instances. For our experiments, we focused on the specific sections of the reports describing the patient's symptoms; therefore, the analysis refers to anonymized data, as the individuals are completely unidentifiable.

As previously mentioned, our objective is to evaluate the clustering method using two distinct types of inputs: the content of each considered section in its original form, i.e., unstructured text in Italian, and the sets of HPO standardized phenotype names automatically extracted from it.

To represent these textual contents in a way suitable for computational analyses, we used the MPNet model [12] to generate document embeddings. We specifically chose the MPNet model since it has been shown to outperform other state-of-the-art competitors; moreover, it is possible to run it on a laptop computer, thanks to its limited number of parameters (109M). The embeddings obtained with MPNet map each document section into a vector space, preserving semantic relationships and ensuring that documents with similar symptom descriptions are represented by vectors close to each other.

The clustering process was performed using the k-means algorithm [6] applied to the document embeddings. This unsupervised machine learning technique allowed us to group documents based on their semantic similarity, facilitating the identification of meaningful patterns in the dataset. According to preliminary tests, the number of clusters was set to 58, as this value provided a good trade-off between the silhouette score and average cluster size.

Clusters can be used for further statistical and predictive analysis. As an example, we computed the percentage of positive cases in each cluster. Results are available in our GitHub repository [11].

2.1 Phenotype Extraction Pipeline

This section outlines the pipeline we developed for extracting HPO [5] phenotype names from the clinical reports. It is fed with the JSON file containing the sections of the medical report describing the patient's symptoms, written in the natural language (Italian). It processes these files in three distinct steps, ending in the generation of a set containing the phenotypes described in the reports.

Step 1: Translation to English. The texts are translated into English to ensure compatibility with the processing capabilities of LLMs. This step uses the Google Translate API via its Python package. While this step is optional, it is recommended due to the significantly higher performance of LLMs when working with English text compared to other languages [8,14].

Step 2: Extraction of phenotypes. The pipeline uses LLaMa 3.2 [9], an LLM with 70 billion parameters, to extract a JSON-encoded list of phenotypes from the translated description of the patient's symptoms. During this step, which forms the core functionality of the data extraction process, our LLM is asked to remove from the list of symptoms all those that are not referred to the patient and to provide its response in JSON format, reporting the symptom extracted from the text and its possible mapping to HPO.

Step 3: Phenotype standardization. To address the occasional output of synonyms or non-existent phenotypes by LLaMa, the extracted phenotypes are standardized using the HPO ontology [5], which provides a comprehensive set of phenotypes along with their hierarchical relationships.

3 Discussion

We engaged domain experts to assign labels to each cluster to semantically interpret the clusters and compare the two approaches. To this end, we generated the set of the most relevant words for each cluster by computing the average TF-IDF [1] score of the words within it. The labeling procedure involved analyzing these top words and reviewing a sample of documents from each cluster.

To assess the clustering quality, we calculated the silhouette score [10] for each cluster. This metric, ranging from -1 to 1, quantifies how cohesive and well-separated a cluster is, with higher values indicating better quality. Labels and silhouette scores are available in our GitHub repository [11].

Many of the clusters belong to one of these two macro-areas: heart diseases and neurodevelopmental disorders. This is because most of the patients in our cohort present with one of these two issues. The main difference, however, lies in the fact that while heart diseases are often related to specific clinical suspicions (e.g., cardiomyopathies or Brugada syndrome), the descriptions for neurodevelopmental disorders are more nuanced and varied: these encompass a wide range of partially overlapping phenotypes (e.g., psychomotor delay, autism, brain anomalies, intellectual disability, etc.), resulting in lower silhouette scores and clusters that should ideally be grouped into a single container.

It has been observed that a silhouette score greater than 0.3 generally enables the clinical domain of a cluster to be defined. However, according to our experiments, it has not always proven to be a reliable parameter for assessing cluster quality. In some cases, reports were grouped together based on purely syntactic aspects. For example, in a cluster with a silhouette score of 0.35, very different diseases were grouped together simply because they are defined as "disease of" (e.g., Caroli disease, Hirschsprung disease, Pompe disease). This issue deserves attention and suggests further experimentation: the use of bigger or more powerful LLMs for the embedding could be explored, or modification to the preprocessing phase could be considered. Clusters with a silhouette score close to 0 or even with a negative value were also observed. In these clusters, points correspond to highly heterogeneous descriptions. These clusters appear to capture outliers, as they encompass a wide range of phenotypes.

Comparing the text-based clustering to the phenotype-based clustering, the latter appears to perform better, as indicated by the higher average silhouette score. However, it should be noted that the second approach introduces an additional bias beyond the translation: associating phenotypes with their corresponding HPO terms. In some cases, this seems to be a successful strategy, such as with RASopathies, which form a standalone cluster.

4 Conclusion and Future Directions

Although this work is still in the early stages, the results are promising and give many insights for future research steps. First, further predictive analyses can be conducted, such as estimating the likelihood of a patient carrying specific genetic

mutations by assessing the percentage of positive cases across different clusters. Also, as suggested in the discussion, alternative models for embedding need to be tested so as to capture semantic similarities and differences better. Similarly, different clustering strategies (e.g., fuzzy [13], hierarchical [7], neural [4]) could be better suited to represent the semantics of this kind of dataset and deserve to be tested. Finally, we intend to assess the quality of the extraction of HPO concepts by LLaMa to eliminate any potential threats to our conclusions arising from inaccuracies or errors in the data extraction pipeline.

In general, our approach is generalizable and applicable to other clinical datasets and has the potential to offer valuable support to pathologists in decision-making, diagnosis formulation, and in the search for clinically similar reports.

Acknowledgments. This work has been partially funded by PNC - ANTHEM (AdvaNced Technologies for Human-centrEd Medicine) - Grant PNC0000003 - CUP: B53C22006700001 - Spoke 1 - Pilots 1.1 and 1.4.

References

1. TF–IDF. Encyclopedia of Machine Learning, pp. 986–987. Springer US, Boston, MA (2010)
2. Baddour, M., Paquelet, S., Rollier, P., De Tayrac, M., Dameron, O., Labbé, T.: Phenotypes extraction from text: analysis and perspective in the llm era. In: 2024 IEEE 12th International Conference on Intelligent Systems (IS), pp. 1–8. IEEE (2024)
3. Bhattarai, K., et al.: Leveraging GPT-4 for identifying cancer phenotypes in electronic health records: a performance comparison between GPT-4, GPT-3.5-turbo, flan-t5, llama-3-8b, and spacy's rule-based and machine learning-based methods. JAMIA Open **7**(3), ooae060 (2024)
4. Fard, M.M., Thonet, T., Gaussier, É.: Deep k-means: Jointly clustering with k-means and learning representations. Pattern Recognit. Lett. **138**, 185–192 (2020)
5. Gargano, M.A., Matentzoglu, N., Coleman, B., et al.: The human phenotype ontology in 2024: phenotypes around the world. Nucleic Acids Res. **52**(D1), D1333–D1346 (2023)
6. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press (1967)
7. Murtagh, F., Contreras, P.: Algorithms for hierarchical clustering: an overview. WIREs Data Mining Knowl. Discov. **2**(1), 86–97 (2012)
8. Qin, L., et al.: A survey of multilingual large language models. Patterns **6**(1), 101118 (2025)
9. Roumeliotis, K.I., Tselikas, N.D., Nasiopoulos, D.K.: Llama 2: Early adopters' utilization of meta's new open-source pretrained model (2023)
10. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
11. Saletta, M.: https://github.com/Martisal/phenoClustering (2025)

12. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.: MPNet: Masked and permuted pre-training for language understanding. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS (2020)
13. Yang, M.S.: A survey of fuzzy clustering. Math. Comput. Model. **18**(11), 1–16 (1993)
14. Zhang, Z., Zhao, J., Zhang, Q., Gui, T., Huang, X.: Unveiling linguistic regions in large language models. In: Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL, pp. 6228–6247. Association for Computational Linguistics (2024)