Contextual Categorization Enhancement through LLMs Latent-Space

Zineddine Bettouche, Anas Safi, Andreas Fischer
Deggendorf Institute of Technology
Dieter-Görlitz-Platz 1, 94469 Deggendorf
zineddine.bettouche@th-deg.de, anas.safi@stud.th-deg.de, andreas.fischer@th-deg.de

***Abstract*—Managing the semantic quality of the categorization in large textual datasets, such as Wikipedia, presents significant challenges in terms of complexity and cost. In this paper, we propose leveraging transformer models to distill semantic information from texts in the Wikipedia dataset and its associated categories into a latent space. We then explore different approaches based on these encodings to assess and enhance the semantic identity of the categories. Our graphical approach is powered by Convex Hull, while we utilize Hierarchical Navigable Small Worlds (HNSWs) for the hierarchical approach. As a solution to the information loss caused by the dimensionality reduction, we modulate the following mathematical solution: an exponential decay function driven by the Euclidean distances between the high-dimensional encodings of the textual categories. This function represents a filter built around a contextual category and retrieves items with a certain Reconsideration Probability (RP). Retrieving high-RP items serves as a tool for database administrators to improve data groupings by providing recommendations and identifying outliers within a contextual framework.**

***Index Terms*—Natural Language Processing, Contextual Categorization, Large Language Models, BERT, Convex Hull, Hierarchical Navigable Small Worlds, High-dimensional Latent Space, Dimensionality Reduction.**

I. Introduction

In the modern age of abundant textual data, effective categorization poses a significant challenge due to its increasing complexity. As data volumes grow exponentially, traditional methods struggle to handle the task adequately.

Fortunately, advancements in Natural Language Processing (NLP) provide a promising solution. NLP techniques, like word embeddings, encode semantic information, enabling automated and accurate categorization of vast datasets. Long Short-Term Memory (LSTM) networks, a type of recurrent neural network, excel at modeling sequential data and have proven effective in automating categorization tasks.

Transformer models [1], such as the BERT series [2], offer a deep understanding of contextual nuances, enhancing categorization accuracy. Innovative algorithms like Convex Hull [3] and Hierarchical Navigable Small World (HNSW) [4], powered by embeddings, can be employed to check the categorization efficiency by organizing and navigating through high-dimensional spaces.

In our previous works [5], [6], we addressed the contextual clustering of transformer encodings in an unlabeled database of scientific articles. In this study, we investigate the effectiveness of these methodologies—NLP techniques, LLMs, and geometric algorithms—in improving categorization efficiency and accuracy. Through empirical analysis, we aim to demonstrate their efficacy in managing large-scale data categorization challenges.

As for the structure of this paper, Section II offers a background on key topics such as Transformer models, and the NLP techniques employed. Section III cites the related works, setting a foundation and context for the research. Section IV lays out the methodology, detailing the approach, models, and metrics used. Section V presents the experiments, the challenges faced, trade-offs considered, and the results derived from the techniques employed. Finally Section VI concludes the study and sets up future work.

II. Background

This section introduces the background of the techniques used to develop our methodology.

A. Transformer Models: BERT

Transformer models, such as BERT, are pivotal tools in the field of natural language processing. BERT, which stands for Bidirectional Encoder Representations from Transformers, has transformed the way we handle text data. BERT models are pretrained on vast textual corpora and excel at converting text into high-dimensional vectors. They shine at capturing intricate semantic relationships between words and sentences, making them invaluable for various NLP tasks.

B. Convex Hull

Convex hulls are key concepts in computational geometry and mathematics. They represent a closed, convex shape that encloses a set of data points in multi-dimensional space. In our study, we use convex hulls to define boundaries around groups of BERT-encoded articles, creating distinct regions within the latent space. This method aids in exploring how articles are distributed within a specific category and identifying those that are positioned near the boundaries. This enables us to adjust and extend category definitions effectively.

C. Hierarchical Navigable Small Worlds

Hierarchical Navigable Small Worlds (HNSWs) serve as a data structure for approximate similarity searches in high-dimensional spaces. They offer a practical and scalable means to navigate complex, multi-dimensional data while preserving

proximity relationships. In our research, HNSWs are employed to efficiently organize and search BERT-encoded vectors. This facilitates the process of identifying articles that closely resemble a given category within the latent space. The hierarchical nature of HNSWs enhances our ability to adjust and expand categories based on latent similarities.

III. Related Work

In this section, we provide a concise overview of related work, highlighting key contributions that inform our study.

Moas' investigation into "Real-time prediction of Wikipedia articles' quality" [7] examines various assessment methods, automated and manual, highlighting the challenges in manual assessment due to its time-consuming nature and issues like "circular categories." The paper suggests using an Application Programming Interface (API) to address these challenges in analyzing category trees. Another study, "The Use of NLP-Based Text Representation Techniques to Support Requirement Engineering Tasks," [8] discusses levels of Natural Language Processing (NLP). It emphasizes the semantic level for understanding text meaning and supports the adoption of the Vector Space Model [9] for its simplicity in representing articles in high-dimensional spaces.

The paper "Data Mining with Python" [10] by Nielsen highlights NetworkX and DiGraphs' utility in analyzing hierarchical structures like category trees in Wikipedia. These directed graphs accurately represent parent-child relationships, essential in modeling category relationships. Studies on convex hulls, such as Preparata and Hong's computational aspects [11] and Graham's algorithm [3], provide insights into computational methods for points in two and three-dimensional spaces. Understanding convex hull derivation from simple polygons aids in comprehending data distribution in high-dimensional spaces, relevant to machine learning.

The research by Malkov and Yashunin [4] introduces Hierarchical Navigable Small World (HNSW) graphs for kNN search, implemented in this thesis for vector retrieval. The analysis revealed that articles within the same category tend to be the closest neighbors in the constructed HNSW structure.

Johnson et al. [12] proposed a language-agnostic method for categorizing Wikipedia articles, overcoming content quality and geographic variations. Leveraging article links, their approach matches language-dependent methods in performance, extending coverage across Wikipedia languages. Biswas et al. [13] introduced Cat2Type, enhancing entity typing in knowledge graphs (KGs) like DBpedia and Freebase using Wikipedia categories. By utilizing semantic information, Cat2Type surpasses existing methods on real-world datasets. Ostendorff et al. [14] addressed semantic relationship identification between documents, treating it as a pairwise classification task. Employing BERT and XLNet, experiments on Wikipedia article pairs and Wikidata properties validated BERT's effectiveness, suggesting potential for semantic document-based recommender systems.

This overview summarizes findings from various research domains, including article categorization, NLP techniques, dimensionality reduction, graph analysis, convex hulls, and nearest neighbor search, all relevant to this study. Notably, the reviewed research does not directly address the use of transformer-generated encodings in their high-dimensionality (full information), which differentiates the methodology presented in this paper. The paper introduces a novel mathematical function that utilizes these encodings, representing a distinctive contribution to the field.


```
{
 "id": "17279752",
 "text": "Hawthorne Road was a cricket and football ground in Bootle in England.
 "title": "Hawthorne Road"
},
{
 "id": "17279785",
 "text": "\"Nothing Lasts Forever\" is a single by Echo & the Bunnymen which was
 "title": "Nothing Lasts Forever (Echo &amp;amp; the Bunnymen song)"
},
{
 "id": "17279795",
 "text": "KPTY (1330 AM, \"107.3 Hank FM\") is a radio station serving the Water
 "title": "KPTY (AM)"
},
```


Figure 1. Sample of Data Objects in the Wikipedia Dataset

IV. Methodology

This section presents the Wikipedia dataset used in this work and the approaches developed to select articles for context-based reconsideration.

A. Overview of Wikipedia Dumps

The data used is from the Wikipedia dump of November 2020 on Kaggle [15] (plain text). There are 6,144,363 articles in the dataset divided into 605 JSON files. Figure 1 shows a random sample of items in these JSON files. Each item has an id, a text, and a title.

The transformer model processes articles individually, ensuring consistent and undistorted encodings regardless of the number of articles processed and sample size. Therefore, as a proof of concept and for computational feasibility, we randomly selected 10,000 articles to assess the approaches and conduct experiments.

B. Convex Hull

Our second approach involves the construction of a convex hull on the set of vectors representing the category. The vectors of the category tree are mapped onto a 2D plane using UMAP [16]. The reason behind including UMAP is that it is not feasible to construct a convex hull in the 768 dimensions Every article in the dataset undergoes encoding and mapping onto the 2D plane, enabling the determination of whether any article breaches the established convex hull boundary. Convex hulls provide an efficient way to capture the spatial relationships between category vectors when mapped onto a 2D plane. This approach leverages the geometric properties of convex hulls, enabling the visualization and definition of the boundaries of category clusters. By determining whether an article breaches the established convex hull boundary, one can promptly identify articles that do not conform to their designated categories.

C. HNSW

The application of Hierarchical Navigable Small World (HNSW) is used in the third approach to retrieve the nearest neighbors of articles within a category tree. The premise of this approach is that retrieved articles should already share the same category. The presence of articles that are retrieved but do not belong to the category tree should be reconsidered for category addition. We construct an HNSW using the randomly selected articles, alongside category vectors. For each category vector, we retrieve the five nearest neighbors in this HNSW. Articles that appear and are not associated with the category, are flagged for inspection. This approach leverages the structure of the data. Articles within the same category should naturally be closer to each other in the vector space. By flagging articles retrieved by HNSW that do not belong to the category tree, one can promptly identify instances of lack of categorization.

D. Filter built on High-dimensional Latent Space

The high dimensionality of the encodings can be too complex for several techniques, and it is referred to as *the curse of dimensionality* in this case. However, we intend to use these information-rich vectors to our advantage. We design a filter that takes the category articles (as encodings) and its centroid vector. This filter is applied on the articles of the sample. It takes a percentage that represents the Reconsideration Probability (RP).

An article with 100% RP must be reconsidered to be added to the concerned category. We encode the category into latent space and calculate the centroid vector. The category encodings form a cloud around this centroid vector in the latent space. Any non-category vector in the dataset that has a distance to the centroid vector equal to or less than the radius of this category cloud is assigned 100% RP.

We assume that in the sample there must at least one article that must not be in the selected category. We assign this non-category article 0.1% RP (0% RP complicates the mathematical modeling of the filter). In the case that the dataset admin cannot select such an article, we assign 0.1% RP to the article with the farthest encoding from the centroid.

As for the filter function, let d_c represent the distance of the last quarter threshold (75th percentile). This value is often used as a measure of central tendency that is less affected by outliers compared to the mean. Let d_{ea} represent the distance of the examined article. The $RP(d_c, d_{ea})$ should be 100% when d_{ea} is equal to d_c, and approach 0% the greater is d_{ea} compared to d_c. We achieve this is by using an exponential decay function with a horizontal shift. To determine the decay constant k, we need to consider the set of non-category articles and their distances. The median value $median(set(d_{ea}))$ in this set of distances should result in 50% RP. Equations 1 and 2 show the mathematical modeling of this description.

$$k = \frac{-ln(0.5)}{median(set(d_{ea})) - d_c} \tag{1}$$

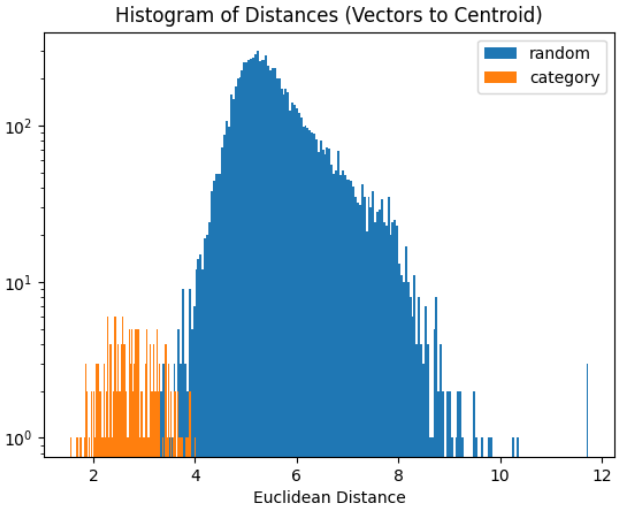

Figure 2. Euclidean Distances between Centroid and Space Vectors

$$RP(d_c, d_{ea}) = 100 * e^{-k*(d_{ea}-d_c)} \tag{2}$$

V. Experiments

This section discusses the experiments done in this paper.

A. Setting up the Vector Space: Encodings of the Category and the Sample Articles

To set the ground for the experiments, the sample of articles and the category articles are all encoded into the vector space. The centroid vector of the category is calculated by averaging its vectors. We calculate then the euclidean distances between this centroid vector and every other vector in this space. Figure 2 shows the results of these calculations. We observe that the distances of the sample articles fall mostly in the interval of distances: $[3.5, 10.5]$ while the category articles have distances to their centroid not exceeding 4, showing an overlap in $[3.5, 4]$. The upcoming experiments shed more light on how to retrieve the closest articles to the centroid (other than the category vectors), and we highlight how these retrieved articles fall distance-wise.

B. Convex Hull: Geometric Boundaries

In this experiment we construct the convex hull with the encodings of the category (Figure 3). The convex hull is placed on the map of the 10,000 articles and the articles that breach it are recorded (Figure 4). The distance-to-centroid of each article breaching the convex hull is recorded and plotted in Figure 5.

The convex hull successfully identified 10 out of the 55 articles with distances smaller than 4 and 33 with distances smaller than 5. This outcome showcases a downside to consider, as the convex hull also selected articles positioned farther away from the centroid than the articles it ignored. This is a form of "blindness" in the method, warranting further exploration and refinement.

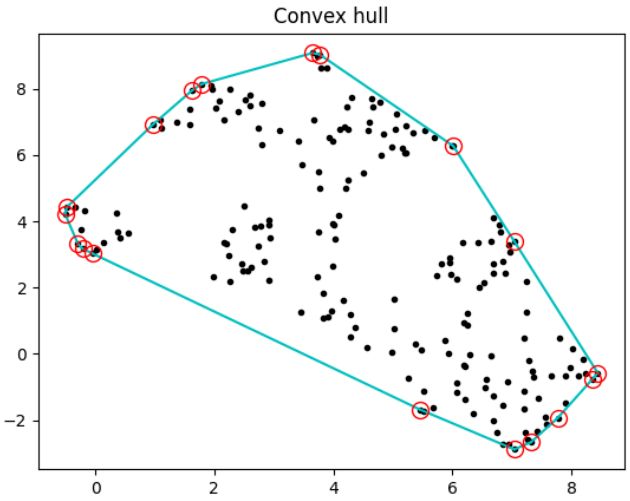

Figure 3. Convex Hull of the Category

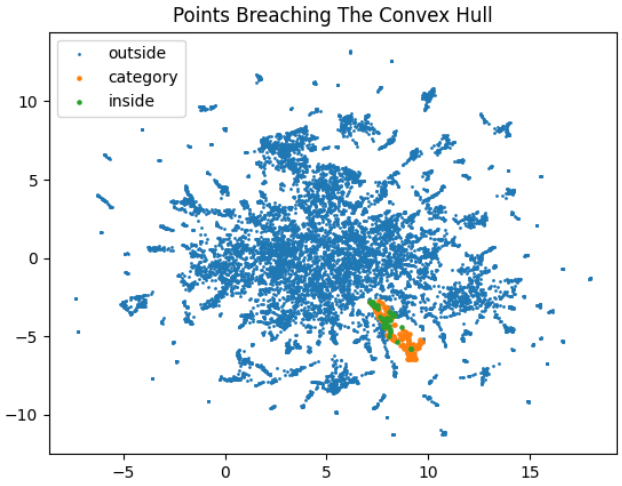

Figure 4. Map of Articles Breaching the Convex Hull

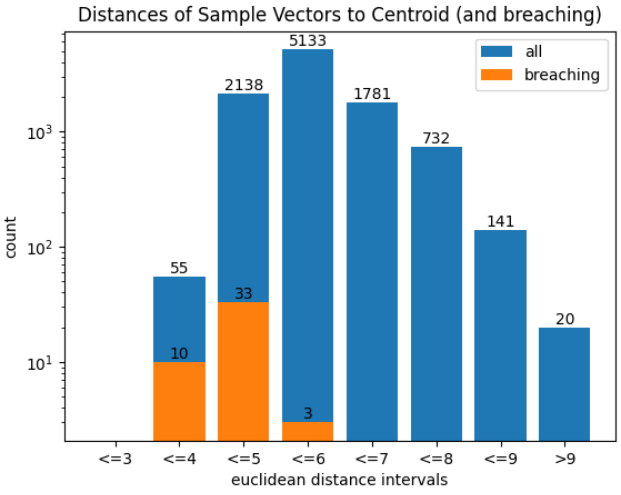

Figure 5. Histogram of Distances of Convex-Hull-Breaching Articles to Category Centroid

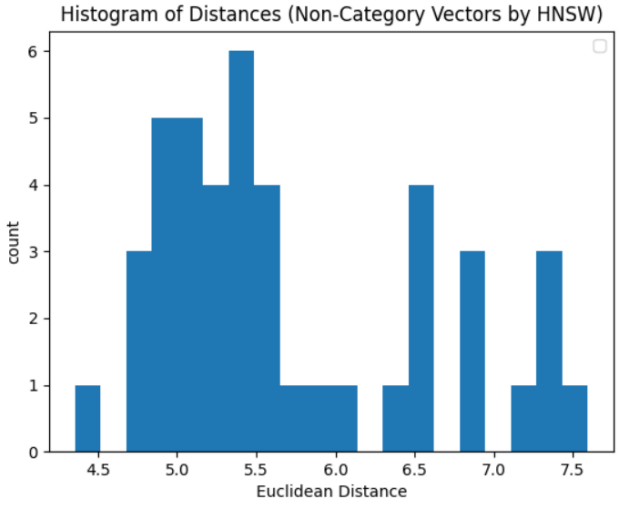

Figure 6. Histogram of Distances of HNSW-Retrieved Articles (non-category) to Category Centroid

C. The Contextual Category in an HNS-World of Articles

An HNSW is built on the setup vectors (sample articles, category articles, and the centroid vector). We retrieve iteratively all of the category vectors from the position of the centroid vector in the HNSW. In this process, we consider the category vectors as a fishnet to retrieve within any other non-category vector that is closer to the centroid vector than the farthest category-vector.

We calculated the distances of the retrieved non-category vectors to the category centroid (Figure 6). We observe that the built HNSW jumped over articles closer to the centroid (only one with $d < 4.5$) and retrieved farther ones (six with $d = 5.4$). This indicates that HNSW follows a pattern similar to the convex hull technique. This similarity arises from the implicit mapping performed by HNSW, resulting in a mapping structure mirroring the loss of information already seen in the convex hull technique.

D. High-Dimensional Latent-Space Filter

As described in IV-D, the 75th percentile distance from the centroid vector to articles of the category d_c is 3.151 (Figure 7). The median of the distances in the non-category vectors $median(set(d_{ea}))$ is 5.447. By numerical application, Equations 1 and 2 are as follows:

$$k = \frac{-ln(0.5)}{median(set(d_{ea})) - d_c} = 0.302$$

$$RP(d_c = 3.151, d_{ea}) = 100 * e^{-0.302*(d_{ea}-3.151)}$$

We calculate the RP values of every articles in the sample (Figure 8). The distances of these articles d_{ea} are in the range of $[3.143, 11.737]$:

- For the minimum value: $RP(3.151, 3.143) = 100.241\%$ (saturated to 100%)

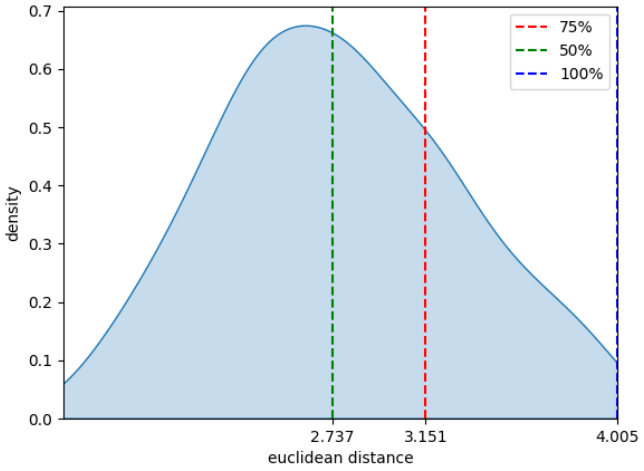

Figure 7. Distances of Category Articles to Category Centroid

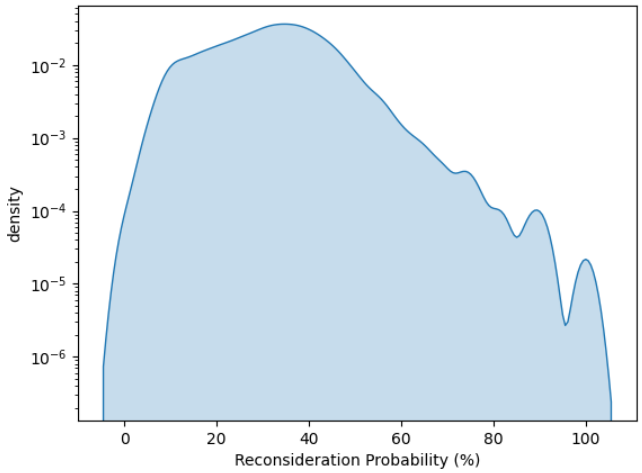

Figure 8. RP of Sample Articles to Category Centroid

- For the median value: $RP(3.151, 5.447) = 49.988\%$
- For the maximum value: $RP(3.151, 11.737) = 7.479\%$

We take the articles with $RP > 75\%$, extract their keywords and plot their wordcloud (Figure 9). The filter retrieved 20 articles with this threshold. Table I shows the breakdown of closest 5 retrieved items. The category "Serbian Films" is similar in context to art topics from the Balkans. This explains the presence of articles about writings from the ex-Yugoslavian countries. To test the variations in centroid vector, we sample the data into 100 samples. The mean distance between the centroid vector and the 80% sample centroid vector is 0.102 with a standard variation of 0.026. This shows an initial stability in the category.

E. Bonus: Hierarchical Vectors Effect on Cluster Cohesion

The presence of hierarchical structures in the form of categories and sub-categories raises the question regarding their utility. One hypothesis worth exploring is whether the incorporation of hierarchical vectors into clusters can reinforce their internal coherence, rendering them more distinct

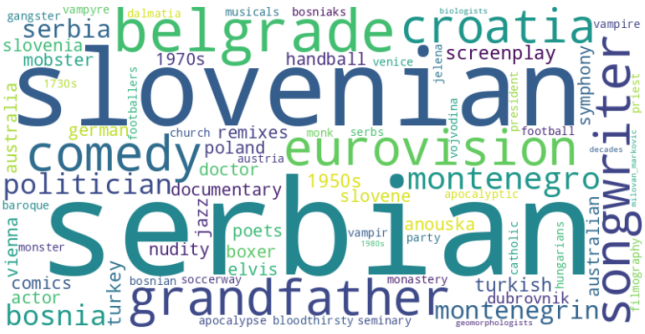

Figure 9. Wordcloud of Articles with $RP > 75\%$

Table I
CLOSEST 5 FILTER-RETRIEVED ARTICLES (CATEGORY: *Serbian Films*)

RP (%)	**Title**	**Keywords**
100.00	Croatia in Eurovision 2006	eurovision serbia montenegro
91.216	Đani Pervan	bosnian musician songwriter
90.993	Vranjic	venice church
90.499	Vinci Vogue Anžlovar	vampyre blog slovenian
89.713	Jovan Cvijić	serbian ethnological historic

from one another. To quantify this distinction, the Silhouette score [17] serves as a reliable metric. To initiate the clustering process, two distinct clusters were created based on articles encoded with BERT, resulting in an initial Silhouette score of 0.23. Subsequently, hierarchical vectors were calculated to represent sub-categories (Figure 10). Each sub-category's vector was computed as the average vector of its constituent encoded articles (further elucidated in the subsequent section). These hierarchical vectors were integrated into the clustering process, and the clustering was re-executed. Notably, the new Silhouette score was improved by 13% (0.26). This shows the impact of incorporating hierarchical vectors, leading to denser clusters.

VI. CONCLUSION AND FUTURE WORK

In conclusion, this paper has addressed the goal of improving the efficiency of contextual categorization within hierarchical structures. This work employed the Wikipedia dumps and its categories, along with BERT models as the latent-

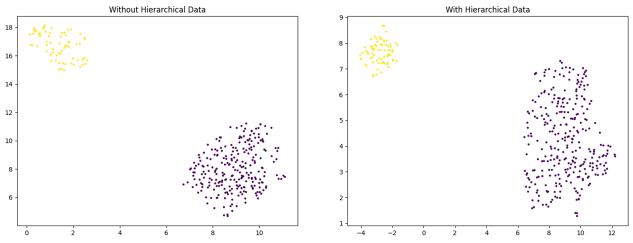

Figure 10. 2D Mappings of two different categories (left: without hierarchical vectors, right: with them).

representation technique. The exploration has focused on the integration of hierarchical vectors and advanced clustering techniques.

The experiments showed the practicality of calculating centroid vectors for category trees. The centroid vector, obtained by averaging all the vectors within a category tree, served as a key item for assessing the proximity of articles to the tree. The relationship between an article's vector and the centroid vector provided a quantitative basis for evaluating the relevance of an article to a specific category tree. This approach allowed for a reconsideration of category labels based on the calculated distances. Alternative techniques such as convex hulls and HNSWs, although explored, exhibited distortions in the results due to the inherent mapping processes involved. To overcome the loss of information and take advantage of the high-dimensionality of the embeddings, we modulated a filter using the exponential decay function that indicates the Reconsideration Probability. The transformer model processes articles individually, ensuring consistent and undistorted encodings regardless of the number of articles processed and sample size. Therefore, the scalability of the exponential decay function (our modulated filter), leveraging transformer embeddings, due to its mathematical nature, offers efficiency gains compared to the scalability challenges typically associated with convex hull and HNSW algorithms. Finally, utilizing hierarchical vectors for subcategories proved valuable, enhancing the representation of the category tree in the latent space. This was evident in the increased Silhouette score, indicating a clearer categorization structure. This utilization is not limited to our case (Wikipedia data), but extends to any form of dataset hosting such hierarchies.

Future research should refine hierarchical vector integration, develop specialized clustering algorithms for complex structures, scale experiments to larger datasets, explore new content categorization tools, assess their impact on platforms like Wikipedia and other databases more scientifically oriented, and enhance categorization systems' precision and utility from a contextual perspective.

ACKNOWLEDGEMENT

This paper has received funding from the state of Bavaria in the context of project SEMIARID, funding no. DIK-2104-0067//DIK0299/01

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Bidirectional encoder representations from transformers," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[3] R. L. Graham and F. F. Yao, "Finding the convex hull of a simple polygon," in *Proceedings of the 19th Annual IEEE Symposium on Foundations of Computer Science*, 1983, pp. 324–331.

[4] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," in *Proceedings of the 34th International Conference on Machine Learning*, 2018, pp. 824–836.

[5] Z. Bettouche and A. Fischer, "Mapping researcher activity based on publication data by means of transformers," in *Proceedings of the Interdisciplinary Conference on Mechanics, Computers and Electrics (ICMECE 2022)*, 2022.

[6] ——, "Topical clustering of unlabeled transformer-encoded researcher activity," *Bavarian Journal of Applied Sciences*, no. 6, pp. 504–525, 2023.

[7] P. M. B. B. L. Moas, "Real-time prediction of wikipedia articles' quality," 2022.

[8] R. Sonbol, G. Rebdawi, and N. Ghneim, "The use of nlp-based text representation techniques to support requirement engineering tasks: A systematic mapping review," 2022.

[9] G. Salton, "Automatic text processing-addison-wesley longman publishing co., inc." 2022.

[10] F. A. Nielsen, "Data mining with python (working draft)," 2017.

[11] F. P. Preparata and S. J. Hong, "Convex hulls of finite sets of points in two and three dimensions," in *Proceedings of the 18th Annual Symposium on Foundations of Computer Science*, 1977, pp. 87–93.

[12] I. Johnson, M. Gerlach, and D. Sáez-Trumper, "Language-agnostic topic classification for wikipedia," in *Companion Proceedings of the Web Conference 2021*, ser. WWW '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 594–601. [Online]. Available: https://doi.org/10.1145/3442442.3452347

[13] R. Biswas, R. Sofronova, H. Sack, and M. Alam, "Cat2type: Wikipedia category embeddings for entity typing in knowledge graphs," in *Proceedings of the 11th Knowledge Capture Conference*, ser. K-CAP '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 81–88. [Online]. Available: https://doi.org/10.1145/3460210.3493575

[14] M. Ostendorff, T. Ruas, M. Schubotz, G. Rehm, and B. Gipp, "Pairwise multi-class document classification for semantic relations between wikipedia articles," in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, ser. JCDL '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 127–136. [Online]. Available: https://doi.org/10.1145/3383583.3398525

[15] DavidShapiro, "Plain text wikipedia dataset 202011," https://www.kaggle.com/datasets/ltcmdrdata/plain-text-wikipedia-202011, 2020, accessed on November 2, 2023.

[16] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection," *arXiv preprint arXiv:1802.03426*, 2018.

[17] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.