LLM Based Approach to Extracting Smells in Slovenian Corpora

Janez Brank
Jožef Stefan Institute
Ljubljana, Slovenia
janez.brank@ijs.si

Inna Novalija
Jožef Stefan Institute
Ljubljana, Slovenia
inna.koval@ijs.si

Dunja Mladenić
Jožef Stefan Institute
Ljubljana, Slovenia
dunja.mladenic@ijs.si

Marko Grobelnik
Jožef Stefan Institute
Ljubljana, Slovenia
marko.grobelnik@ijs.si

Abstract

This paper presents a comparative study of automatic smell detection in Slovenian cultural heritage texts using both keyword-based search and large language model (LLM) inference. We process a portion of the dLib.si corpus from the late 19th and early 20th centuries, analyzing over 1.6 million text segments for olfactory references. The keyword method leverages an expert-curated list of smell terms, while the LLM method applies semantic inference via prompt-engineered queries. We compare the methods in terms of detection density, temporal trends, and agreement overlap. Additionally, we visualize the semantic landscape of extracted smell terms using t-SNE and unsupervised clustering with auto-generated labels. Our findings reveal limited overlap between methods, a shared rise in smell mentions over time, and distinct semantic clusters ranging from industrial to culinary and bodily smells. This study highlights the value of combining symbolic and neural approaches for nuanced sensory mining in digital heritage corpora.

Keywords

LLM, Artificial Intelligence, Cultural Heritage, Text Mining

1 Introduction

Olfactory perception is an essential yet underexplored dimension in the analysis of historical texts, particularly within the cultural heritage domain. Smells, though intangible, play a critical role in shaping memory, atmosphere, and cultural meaning. However, their representation in written sources is often subtle, indirect, or metaphorical. This challenge becomes more pronounced in historical corpora such as 19th- and early 20th-century Slovenian publications, where evolving linguistic practices and cultural norms affect how sensory information is encoded.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2025, Ljubljana, Slovenia
© 2024 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2025.sikdd.5

This paper explores automatic smell detection in Slovenian cultural heritage texts using two complementary strategies: (1) a keyword-based approach derived from an expert-curated list of smell-related expressions and their morphological variants, and (2) large language model (LLM) - based semantic inference using prompt-engineered queries via the Together.ai platform. We process a subset of the dLib.si digital library corpus of Slovenian texts, divided into temporal buckets, and evaluate the performance, overlap, and divergence between the two methods.

To facilitate large-scale analysis, we produce and analyze over 1.6 million document-query pairs, extracting smell mentions, classifying them by agreement type, and visualizing their distributions both temporally and semantically. Our goals are twofold: (i) to quantify the representational density of olfactory references in the corpus, and (ii) to better understand how computational methods can surface subtle cultural patterns that evade traditional keyword search alone.

This work contributes toward a richer modeling of sensory information in digital heritage collections and highlights the value of combining symbolic and neural methods for text mining in the cultural heritage domain.

2 Related Work

Recent years have seen increased interest in the computational modeling of olfactory expressions in historical and cultural texts. A prominent initiative in this space is the Odeuropa project [7], which focused on identifying, curating, and semantically linking smell-related content in European heritage corpora. Large-scale initiatives, such as the Odeuropa project, have produced the European Olfactory Knowledge Graph and tools like the Smell Explorer to trace historical olfactory knowledge across 400 years of European sources [7, 5]. Research on sensory perception in NLP has traditionally focused on the visual and auditory modalities, while olfaction remains relatively underexplored. Annotation frameworks such as the Olfactory Event Frame and guidelines for labeling sources, qualities, and experiences [6] provide structured resources for information extraction from historical and literary corpora. Traditional approaches to olfactory semantics rely on fixed lexicons such as the Dravnieks Atlas [1] and the DREAM challenge descriptors [3]. For morphologically complex and

low-resource languages such as Slovene, monolingual models like SloBERTa [10] and seq-to-seq models like SloT5 [9] demonstrate that tailoring architectures to linguistic structure improves performance over multilingual baselines. A wide range of Slovene corpora underpins these modeling efforts. Gigafida 2.0, a reference corpus of 1.1 billion tokens covering contemporary written Slovene, provides a large-scale foundation for model pretraining and evaluation [4]. For user-generated content, the JANES corpus supplies richly annotated Slovene social media text, including normalization and NER [2]. Unlike prior studies that primarily focus on annotation frameworks, fixed olfactory lexicons, or large-scale multilingual heritage initiatives such as Odeuropa, our work provides the first comparative evaluation of keyword-based and LLM-based smell detection specifically for Slovenian cultural heritage corpora, highlighting the interplay between symbolic coverage and neural semantic inference.

3 Corpora and Preprocessing

For the experiments presented in this paper, we used texts from the Slovenian Digital Library (dLib.si). Initially we downloaded, from the Library's website, all documents from the period 1870–1919 for which OCRed text was available and whose language was marked as Slovene in the metadata there. In terms of content, this covers nearly all books, newspapers, magazines etc. published in Slovene during that period. From this corpus we then randomly selected 7 % of the documents from each year for further processing; thus the selected subset maintains the same distribution over time, genre, etc. as the full corpus. This resulted in a dataset of approx. 366 thousand documents with a total of 105 million words.

4 Methodology

This section outlines the analytical pipeline used to detect, compare, and interpret smell-related expressions in Slovenian cultural heritage texts. Our approach combines large language model inference, keyword-based retrieval, temporal and density statistics, and unsupervised semantic clustering.

4.1 Comparative Evaluation of Detection Methods

In order to identify olfactory expressions, we employed two complementary strategies:

- **LLM-based Extraction:** Each document was split into passages and processed using a LLM.[1] The model returned a list of potential smell-related words or phrases, structured in JSON format. In cases of formatting failure, raw strings or exception messages were recorded.
- **Keyword-Based Search:** A manually curated index of smell-related expressions, including morphologically inflected forms, was used for direct string matching within each passage.[2]

[1]The Llama-3.3-70B-Instruct-Turbo-Free model, accessed via Together.ai.
[2]This index has been kindly provided by Mojca Ramšak and is based on her work on the anthropology of smell [8].

For each passage, we recorded both LLM and keyword results. We classified outcomes into four categories: *LLM Only*, *Keyword Only*, *Both*, or *None*. Additionally, we computed the **Jaccard similarity** J between the two result sets:

$$J(A, B) = |A \cap B| \,/\, |A \cup B|,$$

where A is the set of LLM-based results and B is the set of keyword-based results. This metric enabled quantitative comparison of coverage and intersection across detection methods.

4.2 Temporal Distribution of Smell Mentions

We extracted the year of publication from each document's metadata. For each year, we aggregated:

- Total LLM-detected smell terms
- Total keyword-detected smell terms
- Number of processed queries

These aggregates were used to generate yearly time series, revealing longitudinal patterns in olfactory expression across the corpus. This temporal analysis supports hypotheses about cultural shifts, such as increasing industrial or bodily smell discourse over time.

4.3 Semantic Typology via Clustering of Smell Terms

To explore latent smell categories, we constructed a semantic typology using the following steps:

- **Term Extraction:** We extracted the 500 most frequent smell-related terms from the combined LLM and keyword results.
- **Vectorization:** Terms were embedded using TF-IDF vectors over character-level n-grams (`char_wb` with range 2–4), capturing morphological similarity.
- **Dimensionality Reduction:** The high-dimensional vectors were projected to two dimensions using **t-SNE** (perplexity = 30), yielding a visual semantic landscape.
- **Clustering:** We applied ***k*-means clustering** (with $k = 8$) to the t-SNE coordinates. For each cluster, the top 5 TF-IDF terms were used to generate semantic labels (e.g., *"Herbs & Cooking"*, *"Pharmaceutical Smells"*).
- **Visualization:** The clusters were visualized with color-coded labels and representative terms. Interactive versions were built using `plotly`.

This typology enables data-driven classification of smell discourse and provides interpretable categories for cultural and linguistic analysis.

4.4 Document-Level Smell Density Analysis

To assess the distribution of olfactory content across documents, we computed the *smell density* as the ratio of detected terms to queries per document:

$$\text{Density}_{\text{LLM}} = \frac{\#\text{ LLM terms}}{\#\text{ queries}}$$

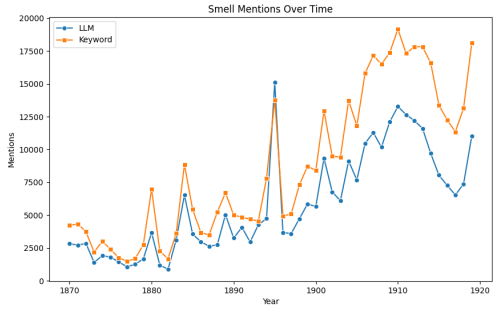

Figure 1: Yearly trends in smell term mentions. Keyword-based detection consistently returns higher frequencies than the LLM, but both show similar growth patterns.

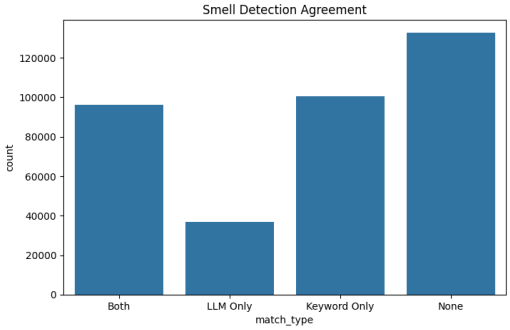

Figure 2: Detection agreement between LLM and keyword methods. Most passages are matched by one method only, with a significant number showing no detection. The overlap ("Both") occurs in fewer than one-third of cases.

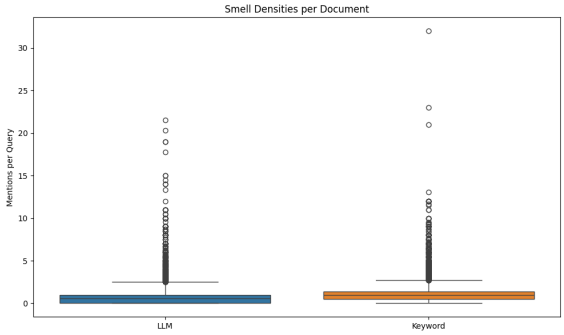

Figure 3: Smell term density per document. While outliers exist for both methods, keyword-based detection generally identifies a higher density of smell references per query.

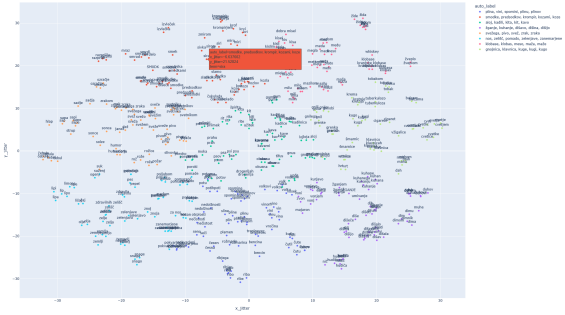

Figure 4: t-SNE semantic landscape of smell terms, clustered by character-level similarity and automatically labeled using top TF-IDF terms per group. The visualization reveals coherent groups such as food, ritual, body, and chemical references.

$$\text{Density}_{\text{Keyword}} = \frac{\#\text{ Keyword terms}}{\#\text{ queries}}$$

This metric enabled identification of smell-rich and smell-sparse texts. Density distributions were visualized using boxplots and descriptive statistics, facilitating selection of representative or outlier texts for deeper qualitative analysis.

5 Evaluation and Results

We evaluated complementary approaches to detecting olfactory references in historical corpora: a keyword-based method and an LLM-based classifier. The results highlight both convergences and divergences in performance across time, document density, and semantic coverage.

Figure 1 shows yearly frequencies of smell-related mentions from 1870 to 1920. While keyword-based detection consistently yields higher absolute counts than the LLM, both methods exhibit similar growth trajectories.

Agreement analysis between the two methods (Figure 2) reveals substantial divergence. Only about one-third of passages are identified by both approaches. A large portion is captured exclusively by the keyword method, while the LLM contributes a smaller but meaningful number of unique detections. A significant subset of passages registers no olfactory detection at all, probably because most documents don't mention smell-related topics in the first place.

Figure 3 illustrates the distribution of smell term density per document. Keyword-based detection generally produces higher densities of references, whereas the LLM outputs are sparser but potentially more semantically filtered. Both distributions exhibit long-tailed outliers, where certain documents contain disproportionately high concentrations of olfactory mentions.

To further analyze lexical diversity, we applied t-SNE to embed and cluster smell-related terms (Figure 4). The resulting semantic landscape reveals coherent groupings that align with cultural domains, including food, ritual, body, and chemical references. These clusters highlight the variety of olfactory expressions and suggest that both methods capture complementary facets of the semantic space. The LLM appears particularly adept at recognizing context-dependent terms, while the keyword method anchors clusters in explicit lexical cues.

Overall, the keyword-based approach provides broader coverage and higher frequencies, but at the cost of noise and overcounting. The LLM method, while more conservative, contributes precision and captures context-sensitive

olfactory references that keywords may overlook. The combination of both thus provides a richer and more balanced representation of olfactory discourse in historical texts.

6 Discussion

Our analysis reveals several key insights into olfactory representations in Slovenian cultural heritage texts and the methodological implications of combining LLM-based and keyword-based detection.

First, both detection strategies show meaningful trends over time, with a noticeable increase in smell-related references around the turn of the 20th century. This may reflect broader urbanization, industrialization, and shifts in public health discourse, which intensified the cultural significance of air quality, hygiene, and olfactory environments.

Second, although keyword-based detection consistently returned more hits, the LLM-based method surfaced a distinct set of semantically inferred mentions. As the agreement analysis shows, only a minority of mentions (∼24 %) were matched by both methods. One possible explanation of this would be if neural inference captures more nuanced or contextually implied smell references, such as metaphorical use ("a whiff of suspicion") or implied odors in narrative scenes.

Third, density analysis suggests that LLMs return more sparse but targeted mentions, while keyword detection produces broader but sometimes noisier coverage. This difference is critical for researchers deciding between high recall and high precision when exploring sensory data in historical texts.

Finally, the t-SNE landscape of smell terms uncovered semantically coherent clusters — e.g., medicinal substances, industrial emissions, festive foods, and bodily decay - and allowed us to generate meaningful auto-labels using top TF-IDF terms. Such visualizations provide a valuable tool for cultural historians to engage with thematic patterns across large-scale textual datasets.

Overall, our findings underscore the value of hybrid approaches to cultural text analysis. By comparing symbolic and neural perspectives, we gain both coverage and subtlety, enabling a deeper reconstruction of sensory worlds encoded in the archives.

7 Conclusion and Future Work

We conducted a dual-method analysis of olfactory references in Slovenian historical texts, revealing how keyword search and LLM-based inference each contribute unique perspectives to sensory data mining. Our results show that while the keyword method offers broad lexical coverage, the LLM can detect more subtle, implied, or metaphorical references often overlooked by surface-level matching.

Furthermore, t-SNE clustering of smell terms revealed rich thematic structures — such as food, medicine, pollution, and ritual — highlighting the semantic complexity of olfactory language.

Together, these results demonstrate the complementary strengths of symbolic and neural approaches for enriching digital humanities research, especially in domains like historical sensory studies where annotation is sparse and vocabulary is diffuse.

Several promising directions remain open for further exploration. First, we plan to expand the dataset to cover all documents in the dLib.si corpus, enabling more robust longitudinal and regional analyses. Second, we aim to improve LLM prompts to better handle nested or narrative contexts, including smells embedded in metaphor, irony, or emotional framing.

Another avenue involves extending the classification of smell mentions into functional categories (e.g., pleasant vs. unpleasant, natural vs. artificial, bodily vs. environmental) using additional LLM-based postprocessing. We also intend to explore multilingual smell detection, comparing Slovene with other Central European languages to study cultural convergence and divergence in olfactory discourse.

Finally, we hope to integrate our smell detection pipeline into public digital heritage platforms, providing curators, historians, and linguists with new tools for sensory exploration of archival materials.

Acknowledgements

This work was supported by the Slovenian Research Agency under the project J7-50233.

References

[1] Andrew Dravnieks. 1992. *Atlas of Odor Character Profiles*. ASTM International, (Feb. 1992). ISBN: 978-0-8031-0456-3. DOI: 10.1520/DS61-EB.

[2] Darja Fišer, Nikola Ljubešić, and Tomaž Erjavec. 2020. The janes project: language resources and tools for slovene user generated content. *Language Resources and Evaluation*, 54, 1, pp. 223–246. Retrieved Aug. 27, 2025 from https://www.jstor.org/stable/48740864.

[3] Andreas Keller et al. 2017. Predicting human olfactory perception from chemical features of odor molecules. *Science*, 355, (Feb. 2017), eaal2014. DOI: 10.1126/science.aal2014.

[4] Simon Krek, Špela Arhar Holdt, Tomaž Erjavec, Jaka Čibej, Andraz Repar, Polona Gantar, Nikola Ljubešić, Iztok Kosem, and Kaja Dobrovoljc. 2020. Gigafida 2.0: the reference corpus of written standard Slovene. eng. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Nicoletta Calzolari et al., editors. European Language Resources Association, Marseille, France, (May 2020), 3340–3345. ISBN: 979-10-95546-34-4. https://aclanthology.org/2020.lrec-1.409/.

[5] P. Lisena, T. Ehrhart, and R. Troncy. European olfactory knowledge graph. Zenodo. DOI: 10.5281/zenodo.10709703.

[6] Stefano Menini, Teresa Paccosi, Serra Sinem Tekiroğlu, and Sara Tonelli. 2023. Scent mining: extracting olfactory events, smell sources and qualities. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Stefania Degaetano-Ortlieb, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz, editors. Association for Computational Linguistics, Dubrovnik, Croatia, (May 2023), 135–140. DOI: 10.18653/v1/2023.latechclfl-1.15.

[7] ODEUROPA Project Consortium. 2021–2023. ODEUROPA: negotiating olfactory and sensory experiences in cultural heritage practice and research. https://odeuropa.eu/. EU Horizon 2020 research and innovation programme, grant agreement No. 101004469. Royal Netherlands Academy of Arts and Sciences (KNAW) Humanities Cluster et al., (2021–2023).

[8] Mojca Ramšak. 2025. *Antropologija vonja*. AMEU-ISH, Ljubljana.

[9] Matej Ulčar and Marko Robnik-Šikonja. 2023. Sequence-to-sequence pretraining for a less-resourced slovenian language. *Frontiers in Artificial Intelligence*, 6.

[10] Matej Ulčar and Marko Robnik-Šikonja. 2021. Sloberta: slovene monolingual large pretrained masked language model. In *SiKDD*.