The MUSTI challenge @ MediaEval 2023 - Multimodal Understanding of Smells in Texts and Images with Zero-shot Evaluation

Ali Hürriyetoğlu[1,*], Inna Novalija[2], Mathias Zinnen[3], Vincent Christlein[3], Pasquale Lisena[4], Stefano Menini[5], Marieke van Erp[1] and Raphael Troncy[4]

[1]KNAW Humanities Cluster, DHLab
[2]Jožef Stefan Institute, Slovenia
[3]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg
[4]EURECOM, Sophia Antipolis, France
[5]Fondazione Bruno Kessler, Trento, Italy

Abstract

We ran the MUSTI challenge the second time after the MUSTI 2022 edition by extending the evaluation with a zero-shot evaluation scenario. This was needed as the first iteration showed us there is a lot of room for improvement and zero-shot performance of the state-of-the-art methods is useful in understanding what available models can predict without any training in a new language. We used the same data from MUSTI 2022 for training and evaluation for MUSTI 2023. Additionally, we prepared a second evaluation scenario, which we call zero-shot, in Slovenian, which was not known by the participants before the evaluation phase started. MUSTI 2023 has attracted many teams and state-of-the-art multimodal systems perform better than the systems proposed in MUSTI 2022.

1. Introduction

The manner in which humans engage with smell is a prime example of intangible cultural heritage: the way smells are created, in what situations they are used, but also how they are appreciated are highly culturally dependent. By engaging with expressions of smells in texts and images across multiple genres and multiple languages over a longer period of time, we can gain more insights into how smells have affected human interactions through time.

While smell is of vital importance in our day-to-day lives, little attention has been paid to it within the natural language processing and computer vision communities. While there are some lexicons focused on smell, the Odeuropa text benchmark dataset is the first multilingual, cross-domain text dataset focused on smell references [1]. Similarly, for computer vision, no prior datasets existed until the ODOR challenge dataset was created by members of this task [2]. In the Multimodal Understanding of Smells in Texts and Images (MUSTI) challenge, we bring these modalities together, inviting the research community to explore parallels and complementarities in the way smells are described and depicted in different modalities.

The MUSTI challenge at MediaEval 2023 aims to collect information about smell from digital multilingual text and image collections between the 16th to 20th centuries. More precisely,

MediaEval'23: Multimedia Evaluation Workshop, February 1–2, 2024, Amsterdam, The Netherlands and Online
*Corresponding author.
†These authors contributed equally.
ali.hurriyetoglu@dh.huc.knaw.nl (A. Hürriyetoğlu); inna.koval@ijs.si (I. Novalija); mathias.zinnen@fau.de (M. Zinnen); vincent.christlein@fau.de (V. Christlein); pasquale.lisena@eurecom.fr (P. Lisena); menini@fbk.eu (S. Menini)

MUSTI studies how different smells are referenced in modalities using a corpus of historical multilingual texts and images. For example, what smell references can be identified in a text and what smell sources and/or olfactory gestures can be recognized in an image?

This paper is for the second edition of MUSTI. The first edition in 2022 observed that achieving a good baseline for the task is feasible. One participant submission validated the task by obtaining reasonable performance [3, 4]. However, there remains significant room for improvement in terms of classification performance. Furthermore, the quest for insight has not yet been addressed thoroughly. Additionally, MUSTI 2023 extends the 2022 protocol by adding a zero-shot evaluation setting.

2. Motivation and Background

To fully make sense of digital (heritage) collections, it is necessary to go beyond an ocular-centric approach and to engage with their olfactory dimension as well, as these offer a powerful and direct entry to our emotions and memories. With the MUSTI task, we aim to accelerate the understanding of olfactory references in English, Dutch, French, German, Italian, and Slovene texts and images as well as the connections between these modalities. As recent and ongoing exhibitions at Mauritshuis in The Hague, Netherlands, Museum Ulm in Ulm, Germany, and the Prado Museum in Madrid, Spain demonstrate, museums and galleries are keen to enrich museum visits with olfactory components – either for a more immersive experience or to create a more inclusive experience for differently abled museum visitors such as those with a visual impairment. Reinterpreting historical scents is attracting attention from various research disciplines (Huber et al., 2022) and leading to interesting collaborations with perfume makers, for example, the Scent of the Golden Age candle was developed after a recipe by Constantijn Huygens in a collaboration between historians and a perfume maker. To ensure that such enrichments are grounded in historically correct contexts, language and computer vision technologies can help to find olfactory relevant examples in digitized historical collections and related sources.

With this task, we aim to investigate: *i)* What does it mean for a text and an image to be related in terms of smell? *ii)* Do different text and image genres reference smell differently? *iii)* Do different languages reference smell differently? *iv)* How do references to smell in texts and images change over time? *v)* How do relationships between smell references in texts and images change over time?

3. Task description

Smell is an underrepresented dimension of many multimedia analysis and representation tasks. MUSTI aims to further the understanding of textual descriptions and visual depictions of smells and smelling in historical texts and images. In this shared task, participants are provided with multilingual texts (English, Dutch, German, French, Italian, and Slovene) and images, from the 16th to the 20th century, that pertain to smell in different ways. The images and the texts have been selected because they contain depictions (images) and descriptions (text) of objects that are known to reference smell. The goal of the task is to detect references to depictions (objects such as flowers or animals in an image) and descriptions (texts) of objects that are known to evoke smells in texts and images and to connect these smell references across these two modalities. We formulate the challenge in the following subtasks that could be tackled independently from each other:

Subtask 1: Task participants are invited to develop language and image recognition technologies to predict whether a text passage and an image contain references to the same smell source or not. This task can therefore be cast as a binary classification problem.

Subtask 2: [Optional] The participants are also asked to identify what is (are) the common smell source(s) between the text passages and the images. The detection of the smell source includes detecting the object or place that has a specific smell, or that produces an odour (e. g. plant, animal, perfume, human). In other words, the smell source is the entity or phenomenon that a perceiver experiences with his or her senses. This sub-task can therefore be cast as a multi-label classification problem.

Subtask 3: [Optional] For this subtask we include a new evaluation setting, with test data that consists of image and text pairs in languages that are not provided in the training setting. The training data is available in English, French, German, and Italian and the test data is in all these four languages and two additional languages, which are Dutch and Slovene. We refer to this subtask as a zero-shot evaluation setting.

4. Target groups and Recruiting participants

Due to the growing interest in sensory mining (e. g. 1st International Workshop on Multisensory Data and Knowledge (MDK) @ LDK 2021 and 2nd International Workshop on Multisensory Data and Knowledge (MDK) @ theWebConf 2023) and multimodal information processing (e. g. 1st International Workshop on Multimodal Understanding for the Web and Social Media (MUWS), co-located with The WebConf (WWW) 2022 in different research disciplines. Although participation was limited in MUSTI 2022, we consider MUSTI 2023 to be an opportunity to get in early and establish a leading position on this problem. Community outreach has already started in 2022 and with the execution of a communication plan to enhance the likelihood of reaching a broad community that could propose solutions to the problem we proposed in 2023. The Computer Vision ODOR challenge that we organised as a part of ICPR2022, demonstrates the research community's interest in taking on the previously unaddressed topic of smell. As the task proposers are members of the language technology, computer vision, cultural heritage, digital humanities and semantic web communities, they will publicize the task in their communities via the appropriate mailing lists, social media channels such as Twitter/X and Mastodon, and via upcoming presentations at the Language Resources and Evaluation Conference, the Digital Humanities/Artificial Intelligence Seminar, the European Semantic Web Conference, DHBenelux, The Web Conference, and the Digital Humanities Conference. Furthermore, the Odeuropa Network (consisting of >150 members), the project mailing list, and other communication channels have a wide reach. Finally, we have collected a list of scholars and research groups that work at the intersection of vision and language processing in the first edition of MUSTI in 2022. We will expand this list and invite these people to participate in MUSTI 2023. The MUSTI task also provides an excellent use case for students to hone their multimodal and creative problem-solving skills. We will therefore also advertise the challenge at relevant outlets such as the International Semantic Web Summer School and the EURECOM Machine Learning and Intelligent System (MALIS) course.

By splitting up the task into two stages (first binary classification, then multi-class classification) we aim to reduce the barrier to participation. Furthermore, the team will make available baseline smell reference recognition software for texts and images that the participants can build on.

Most researchers have already very busy agendas thus we aim to make the task attractive to interested parties by providing tools to get going more easily. Furthermore, we will actively

target students and early-career researchers as well as industry to cast a wide net. The potential application domains of the task help here.

The Odeuropa project has created smell reference benchmark datasets for texts and images that will be utilised [1, 2].

5. Data

The MUSTI 2023 dataset consists of copyright-free texts and partly copyrighted images that can be downloaded and submitted by the participants using the URLs we provide. We offer texts in English, Dutch, French, German, Italian, and Slovene (zero-shot scenario) that participants are to match to the images. The texts are selected from open repositories such as Project Gutenberg, Europeana, Royal Society Corpus, Deutsches Text Arxiv, Gallica, Wikisource and Liber Liber. The images are selected from different archives such as RKD, Bildindex der Kunst und Architektur, Museum Boijmans, Ashmolean Museum Oxford, and Plateforme Ouverte du Patrimoine. The images are annotated with 169 categories of smell objects and gestures such as flowers, food, animals, sniffing and holding the nose. The object categories are organised in a two-level taxonomy. The Odeuropa text and image benchmark datasets are available as training data to the participants. The image dataset consists of 4,696 images with 36,663 associated object annotations, 600 gesture annotations, and image=level meta-data. We also provide the output of a text processing system we have developed to identify text snippets that contain smell references. The systems of the participants are evaluated on a held-out dataset of roughly 1,200 images with associated texts in the four languages.

Figure 1 provides an example of mapping images with Slovenian text (text translation: "The stem is round and smooth, and the leaves are lanceolate and bright green. Lily's flowers are large, pure white, and smell very nice. Each flower has six petals, which are curved back at the top. Lily means purity and innocence.") The Slovenian example presents a description of the Lily flower from the journal "Teacher's Mate" published in 1862.

6. Evaluation

Task runs are evaluated against a gold standard consisting of image-text pairs. For the evaluation, we use multiple statistics as each provides a slightly different perspective on the results. The code and models of the baselines are available at . The subtasks are evaluated using the following metrics:

Subtask 1: Predicting whether an image and a text passage evoke the same smell source or not. This subtask is evaluated using precision, recall and F_1-score. As multiple text passages in different languages can be linked to the same image, we employ multiple linking scorers such as CEAF and BLANC to measure the performance across different smell reference chains.

Subtask 2: Identifying the common smell source(s) between the text passages and the images. For this subtask, precision, recall and F_1-score are employed, as well as more fine-grained evaluation methods such as RUFES, which can accommodate multi-level taxonomies.

Subtask 3: Zero-shot evaluation setting. The evaluation for this subtask is the same as subtasks 1 and 2. The only difference is that no training data was provided for this subtask.

Steblo ima okroglo in gladko , perje pa suličasto in svitlo zeleno . Liliin cvetje velik , čisto bel , in prav lepo diši . Vsaki cvet ima šest listikov , kteri so na verhu nazaj zakrivljeni . Lilija pomeni čistost in nedolžnost .

Figure 1: Example from Slovenian data: image and mapped text snapshot.

7. Related Work

To the best of our knowledge, the task of predicting whether an image and a text evoke the same smell has not been tackled prior to the previous MUSTI challenge [3]. However, some closely related tasks about text-image alignment are established in literature: In visual question answering (VQA), the aim is to develop systems capable of reasoning about visual information in order to answer textual questions posed to the systems [5]. Based on existing datasets like COCO [6] or Visual Genome [7], various datasets and benchmarks have been proposed since the mid-2010s to train and evaluate VQA algorithms [8, 9, 10, 11].

Another closely related strand of research is vision-language pretraining (VLP) where multimodal language and vision models are pre-trained on large amounts of image-caption pairs to learn an embedding space shared between visual and textual embeddings. Models pre-trained in this manner exhibit strong generalization capabilities when fine-tuned and applied to their respective downstream task. The most influental VLP algorithm is CLIP [] with numerous applications such as multimodal object detection [12, 13], image retrieval, artwork classification [14], or captioning [15, 16].

Even closer to the MUSTI objective is the task of visual entailment (VE), introduced by Xie et al. [17, 18] together with their SNLI-VE dataset which provides the default benchmark for the task. Given an image-sentence pair, the aim of VE is to predict whether the image semantically entails the text. VE algorithms are thus required to develop a semantic understanding of both images and texts and relate them to each other. Recent algorithms like OFA [19] or PromptTuning [20] achieve accuracies of over 90% at the SNLI-VE benchmark, suggesting that a more difficult benchmark might be beneficial. Given that in MUSTI, logical entailment is replaced with smell entailment, the MUSTI objective could be framed as *olfactory entailment* as opposed to VE.

References

[1] S. Menini, T. Paccosi, S. Tonelli, M. Van Erp, I. Leemans, P. Lisena, R. Troncy, W. Tullett, A. Hürriyetoğlu, G. Dijkstra, F. Gordijn, E. Jürgens, J. Koopman, A. Ouwerkerk, S. Steen, I. Novalija, J. Brank, D. Mladenic, A. Zidar, A multilingual benchmark to capture olfactory situations over time, in: N. Tahmasebi, S. Montariol, A. Kutuzov, S. Hengchen, H. Dubossarsky, L. Borin (Eds.), Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 1–10. URL: https://aclanthology.org/2022.lchange-1.1. doi:10.18653/v1/2022.lchange-1.1.

[2] M. Zinnen, P. Madhu, R. Kosti, P. Bell, A. Maier, V. Christlein, Odor: The icpr2022 odeuropa challenge on olfactory object recognition, in: 2022 26th International Conference on Pattern Recognition (ICPR), IEEE, 2022, pp. 4989–4994.

[3] A. Hürriyetoglu, T. Paccosi, S. Menini, M. Zinnen, P. Lisena, K. Akdemir, R. Troncy, M. van Erp, MUSTI - multimodal understanding of smells in texts and images at mediaeval 2022, in: S. Hicks, A. G. S. de Herrera, J. Langguth, A. Lommatzsch, S. Andreadis, M. Dao, P. Martin, A. Hürriyetoglu, V. Thambawita, T. S. Nordmo, R. Vuillemot, M. A. Larson (Eds.), Working Notes Proceedings of the MediaEval 2022 Workshop, Bergen, Norway and Online, 12-13 January 2023, volume 3583 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3583/paper50.pdf.

[4] K. Akdemir, A. Hürriyetoglu, R. Troncy, T. Paccosi, S. Menini, M. Zinnen, V. Christlein, Multimodal and multilingual understanding of smells using vilbert and muniter, in: S. Hicks, A. G. S. de Herrera, J. Langguth, A. Lommatzsch, S. Andreadis, M. Dao, P. Martin, A. Hürriyetoglu, V. Thambawita, T. S. Nordmo, R. Vuillemot, M. A. Larson (Eds.), Working Notes Proceedings of the MediaEval 2022 Workshop, Bergen, Norway and Online, 12-13 January 2023, volume 3583 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3583/paper36.pdf.

[5] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, A. Van Den Hengel, Visual question answering: A survey of methods and datasets, Computer Vision and Image Understanding 163 (2017) 21–40.

[6] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.

[7] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, International journal of computer vision 123 (2017) 32–73.

[8] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, Vqa: Visual question answering, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2425–2433.

[9] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, Making the v in vqa matter: Elevating the role of image understanding in visual question answering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6904–6913.

[10] Y. Zhu, O. Groth, M. Bernstein, L. Fei-Fei, Visual7w: Grounded question answering in images, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4995–5004.

[11] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, R. Girshick, Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2901–2910.

[12] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, et al., Grounded language-image pre-training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10965–10975.

[13] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al., Grounding dino: Marrying dino with grounded pre-training for open-set object detection, arXiv preprint arXiv:2303.05499 (2023).

[14] M. V. Conde, K. Turgutlu, Clip-art: Contrastive pre-training for fine-grained art classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3956–3960.

[15] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International Conference on Machine Learning, PMLR,

2022, pp. 12888–12900.
[16] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, arXiv preprint arXiv:2301.12597 (2023).
[17] N. Xie, F. Lai, D. Doran, A. Kadav, Visual entailment task for visually-grounded language learning, arXiv preprint arXiv:1811.10582 (2018).
[18] N. Xie, F. Lai, D. Doran, A. Kadav, Visual entailment: A novel task for fine-grained image understanding, arXiv preprint arXiv:1901.06706 (2019).
[19] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, H. Yang, Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, in: International Conference on Machine Learning, PMLR, 2022, pp. 23318–23340.
[20] H. Yang, J. Lin, A. Yang, P. Wang, C. Zhou, H. Yang, Prompt tuning for generative multimodal pretrained models, arXiv preprint arXiv:2208.02532 (2022).