# Multimodal and Multilingual Olfactory Matching based on Contrastive Learning

Sergio Esteban-Romero[1,*], Iván Martín-Fernández[1], Jaime Bellver-Soler[1], Manuel Gil-Martín[1] and Fernando Fernández-Martínez[1]

[1]*Grupo de Tecnología del Habla y Aprendizaje Automático (THAU Group), Information Processing and Telecommunications Center, E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid (UPM)*

Abstract

This paper introduces an innovative approach to the multimodal smell identification task, using CLIP-based solutions employing Vision Transformers (ViT) as image processors and language-specific text encoders. The proposed method addresses the question of whether image-text pairs convey similar olfactory experiences by aligning them in a shared embedding space. A notable consideration in our study is the challenge posed by class imbalance, where certain olfactory experiences have a more significant representation. Hence, this paper describes a supervised methodology during the training of the CLIP-based model, enhancing positive olfactory relationships while mitigating them otherwise. Additionally, we have also explored different data balancing procedures aimed at preserving the original distribution between languages. One of our proposed approaches has demonstrated enhanced accuracy compared to the top-performing result reported in the past 2022 MUSTI challenge edition.

## 1. Introduction and Related Work

In the evolving landscape of artificial intelligence (AI) algorithms, the sensory dimension of smell has been left out of the picture compared to the advances in computer vision, natural language processing, and audio recognition. However, from all that media sources, it is possible to identify which are the olfactory elements present in them. The integration of smell in the digital landscape presents a set of challenges given its ability to be captured indirectly.

Those efforts resulted in challenges like the MUSTI task of MediaEval2022 [1] and MediaEval2023 [2]. In particular, it is focused on the development of systems to determine whether image-text pairs evoke the same smell or not but also to effectively identify which smell sources are present in the images.

Also, best approaches related to the task are based on the use of state-of-the-art models to explore images and text separately to later perform a visual entailment task, as presented by Akdemir et al. [3]. Another relevant solution comes from Shao et al. [4] where Yolov5 is used to extract features that will be encoded alongside the textual passage using a multilingual model to finally feed a classifier.

In this paper, we present a solution based on the creation of language-specific CLIP [5] models using different text encoders for each of the four available languages, and also a combination of

*MediaEval'23: Multimedia Evaluation Workshop, February 1–2, 2024, Amsterdam, The Netherlands and Online*

*Corresponding author.

†These authors contributed equally.

sergio.estebanro@upm.es (S. Esteban-Romero); ivan.martinf@upm.es (I. Martín-Fernández); jaime.bellver@upm.es (J. Bellver-Soler); manuel.gilmartin@upm.es (M. Gil-Martín); fernando.fernandezm@upm.es (F. Fernández-Martínez)

0009-0008-6336-7877 (S. Esteban-Romero); 0009-0004-2769-9752 (I. Martín-Fernández); 0009-0006-7973-4913 (J. Bellver-Soler); 0000-0002-4285-6224 (M. Gil-Martín); 0000-0003-3877-0089 (F. Fernández-Martínez)

them training a simple neural network to perform a linear regression task. Our work is focused on Subtask 1, predicting whether a text passage and an image evoke the same smell source.

## 2. Approach

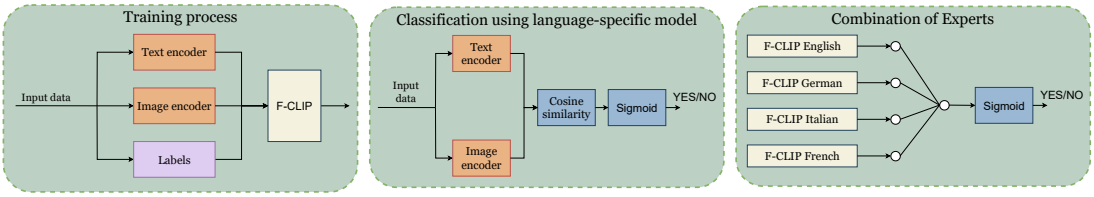


**Figure 1:** Overview of the proposed solutions.

The original implementation of CLIP [5] is based on the premise that for a given image-text pair, both text and image encoders ideally produce identical representations within a shared embedding space. However, common standard implementations of the framework are unsuitable for our specific scenario, since for negative smell relationships, we expect different representations even when they depict similar overall semantic concepts. To extend the CLIP contrastive approach to encompass negative pairs alongside positive ones, we modify the loss function to accommodate both similar (positive) and dissimilar (negative) examples. Consequently, during training, our loss function aims to bring together model representations of positive pairs, fostering similar embeddings, while simultaneously driving apart representations of negative pairs, encouraging dissimilar embeddings. This incorporation of both positive and negative pairs allows the model to discern between relevant and irrelevant image-text pairs, thereby improving its capacity to comprehend and differentiate semantic content.

For the models used for training, we used the ViT[1] checkpoint for all languages for the vision part. Regarding text encoders, those used are English MPNET[2], French Camembert[3], Italian BERT[4] and German BERT[5]. The checkpoints are available at *huggingface*.

Since the number of examples to train using the original challenge dataset is low and in combination with class imbalance, three different experimental setups, described in Table 1, were considered. With these additional experiments, our aim was to obtain a more general and unbiased model that can generalize adequately, but also being potentially more capable of detecting positive cases.

### 2.1. Training language-specific CLIP models

Every language-specific CLIP model is trained by using image-text pair inputs for their corresponding encoders to obtain textual and visual embeddings. Next, we calculate the cosine similarity and utilize it as input for a sigmoid function, ensuring alignment of the resulting output with the specific format demanded by our task. Finally, our approach adopts the Binary Cross-Entropy (BCE) loss.

Note that the usage of different text encoders yields language-dependent visual classifiers. As a result, a projection layer is included after obtaining the embeddings, normalizing the input

[1]https://huggingface.co/google/vit-base-patch16-224-in21k
[2]https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[3]https://huggingface.co/dangvantuan/sentence-camembert-large
[4]https://huggingface.co/dbmdz/bert-base-italian-uncased
[5]https://huggingface.co/bert-base-german-dbmdz-uncased

size while effectively reducing the complexity of the problem. In particular, the final shared space is restricted to a size of 256.

Despite utilizing language-specific text encoders, data from all languages is used when training each fine-tuned CLIP. Although we aimed to develop language-specialized models, we thought it would also be beneficial to learn from others because some words or expressions do not vary much between languages and also because there are few samples for some cases.

### 2.2. Combination of Experts (CoE)

When training every pair of language-specific text and image encoders is performed, we hypothesized that their combination might enhance overall performance. Therefore, to benefit from their distinct expertise, cosine similarities are computed for each pair of examples using language-specific models from all languages. The values obtained are used as input to train a simple neural network with a regression layer, producing a single output that represents the probability of belonging to the positive class. Finally, a threshold is applied to obtain the final classification based on the predictions of the model.

## 3. Results and Discussion

To evaluate our models, we have followed a 5-fold cross-validation scheme. In addition, a fixed reduced test set particular for each of the experimental setups considered was used. A description of them is reported in Table 1. The metrics used are f1 macro, since it is the one used in the challenge to evaluate models, and also the area under the curve (AUC) in combination with receiver operating characteristic (ROC) curves to define the best possible threshold in class imbalance scenarios. For doing so, we used the scikit-learn ROC curve implementation to select it among those proposed. To obtain it, the geometric mean of sensitivity and specificity of each proposal is calculated and the threshold maximizing it is selected.

Regarding the experimental setup with the original challenge dataset, corresponding to unbalanced negative (Unbal. Neg.), a noticeable bias could be observed in the mean of cosine similarities (Mean Cos. Sim.). Consequently, models trained using such data are expected to be biased too. Following the threshold selection procedure defined previously, 0.3 is approximately the best to be used. If we apply it in the experiments carried out under our evaluation scheme, the best performance is obtained for the language-specific model using the French text encoder with a **0.7031**. It also achieves a **0.6342** in the challenge test dataset. Furthermore, the CoE reports a macro-f1 value of **0.4648** in Table 2, showing that the performance is considerably lower in contrast to using one model independently.

For the balanced setup, considering the mean value of the cosine similarities, we can conclude that our models are now less biased than when all available data are used. Looking again at the results in Table 1 and in Table 2, the language-specific French model is the best, obtaining values of **0.7253** and **0.6401**, respectively. In particular, the latter surpasses the best result reported by Akdemir et al. on [3] which is **0.6176**. However, the CoE solution achieves here a macro-f1 of **0.4443** on the challenge test dataset. Finally, if we again compute the best possible threshold to be applied over the probabilities from the models, a value of 0.5 is obtained.

For the case with class imbalance towards the positive class, represented as *Unbal. Pos*, the mean value of cosine similarities suggests that our models are slightly biased. This also highlights that data augmentation processes may be required to train a model with these specific biases, as we thought that removing more examples to enhance a larger bias would lead to low-performance models. In this case, best performing model is obtained using the German

**Table 1**
Distribution of examples for each of the different experimental setups considered with the mean value of the cosine similarity for each image-text pair on each test set. Additionally, f1 macro score obtained under 5-fold cross-validation procedure for each model trained is presented.

| Exp. setup | Total | Pos. | Neg. | Test | Mean Cos.Sim. | English | French | German | Italian | CoE |
|---|---|---|---|---|---|---|---|---|---|---|
| Unbal. Neg. | 2,374 | 593 | 1,781 | 356 (15%) | -0.64±0.05 | 0.6532 | **0.7031** | 0.6814 | 0.6311 | 0.6226 |
| Balanced | 1,218 | 593 | 625 | 122 (10%) | -0.02±0.06 | 0.6889 | **0.7253** | 0.7015 | 0.6504 | 0.6490 |
| Unbal. Pos. | 994 | 593 | 401 | 146 (15%) | 0.26±0.10 | 0.7043 | 0.6873 | **0.7125** | 0.6542 | 0.5950 |

**Table 2**
Results from runs on the challenge test dataset, using models trained on unbalanced negative (Unbal Neg.), unbalanced positive (Unbal Pos.), and balanced (Bal.) datasets.

| Exp. setup | English | French | German | Italian | CoE |
|---|---|---|---|---|---|
| Unbal Neg, | **0.6157** | 0.6342 | 0.5903 | **0.5834** | **0.4648** |
| Balanced | 0.5789 | **0.6401** | **0.6284** | 0.5234 | 0.4443 |
| Unbal. Pos. | 0.5279 | 0.4538 | 0.6193 | 0.5399 | 0.4363 |

text encoder. However, attending to Table 2, the CoE result was evaluated on the challenge test dataset and obtained a score of **0.4363**. Regarding optimal thresholds, in this case values are obtained around 0.55 and 0.6, the latter being selected for the results presented.

## 4. Conclusions

In this paper, we propose a method to simultaneously adapt text and image encoders. When the training process is finished and they are combined, a scent embedding space is obtained, where for image-text pairs with olfactory relationships, their individual representations will be similar and different otherwise. The effectiveness of a single pair of text-image encoders achieves better overall performance compared to the combination of the output of all experts in different languages for this specific problem. Moreover, balancing the dataset has been proven as an effective technique for better generalization. It is also illustrated in Table 1 how after removing almost half of the examples from the original dataset to balance it, some improvements in the f1 score can be observed. This is also highlighted in Table 2 since the best result is obtained when using the balanced setup. In particular, the language-specific French model shows better overall performance, possibly as a result of having used a more powerful text encoder compared to the rest. However, more exploration is required to benefit from the expertise of each language-specific model, as they are proven to work independently.

## Acknowledgments

S.E.-R.'s research was supported by the Spanish Ministry of Education (FPI grant PRE2022-105516). This work was funded by Project ASTOUND (101071191 — HORIZON-EIC-2021-PATHFINDERCHALLENGES-01) of the European Commission and by the Spanish Ministry of Science and Innovation through the projects GOMINOLA (PID2020-118112RB-C22) and BeWord (PID2021-126061OB-C43), funded by MCIN/AEI/ 10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR".

## References


[1] 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.

[2] A. Hürriyetoglu, I. Novalija, M. Zinnen, V. Christlein, P. Lisena, S. Menini, M. van Erp, R. Troncy, The MUSTI challenge @ MediaEval 2023 - multimodal understanding of smells in texts and images with zero-shot evaluation, in: Working Notes Proceedings of the MediaEval 2023 Workshop, Amsterdam, the Netherlands and Online, 1-2 February 2024, 2023.

[3] 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.

[4] Y. Shao, Y. Zhang, W. Wan, J. Li, J. Sun, Multilingual text-image olfactory object matching based on object detection, 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/paper15.pdf.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, CoRR abs/2103.00020 (2021). URL: https://arxiv.org/abs/2103.00020. arXiv:2103.00020.