MUSTI - Multimodal Understanding of Smells in Texts and Images Using CLIP

P. Mirunalini[1,*], V.Sanjhay[1], S Rohitram[1] and M.Rohith[1]

[1]*Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai - 603110, Tamil Nadu, India*

Abstract

This study, titled the Multimodal understanding of Smells in Texts and Images (MUSTI) task, aims to explore the relationship between textual descriptions and visual depictions of smells. Using machine learning techniques, specifically leveraging the CLIP model and tokenization methods, this research extracts features from both text and images to analyze and correlate olfactory elements.The approach involves a model that computes similarity between textual descriptions and images based on their scent-evoking content. This model engages in classification tasks, determining whether a given text-image pair shares a common smell source (1 for positive correlation, 0 for negative correlation). By calculating similarity scores between text and image features, it quantifies the degree of correlation based on scent-related content, enabling a nuanced understanding of connections between textual descriptions and visual representations of smells. Trained on a dataset of image-text pairs, the model outputs scores for accuracy, providing a foundation for a comprehensive analysis of scent-related associations within multimedia content.]

Keywords: Deep learning model, Smells, CLIP, BERT

1. Introduction

The understanding of smells, an underrepresented dimension in multimedia analysis, serves as the focal point of the MUSTI (Multimodal Understanding of Smells in Texts and Images) task. This task aims to delve into the intricate relationship between olfactory references found in textual descriptions and visual depictions across different historical periods and languages. This research work outlines the task's primary objectives: MUSTI Classification, requiring participants to predict shared olfactory sources between texts and images as a binary classification problem. Utilizing the CLIP model and tokenization methods, the model extracts features from images and text and computes the cosine similarity between them based on smell-related content.

2. Related Work

In recent years, the study of olfactory dimensions has gained traction, recognizing the significance of smells in memory and emotions [1].Projects like Odeuropa aim to enrich metadata and develop new interfaces like the "scent wheel" in cultural heritage collections [2]. Initiatives, including a diachronic multilingual benchmark and a comprehensive data model, address this gap by capturing and structuring olfactory information across languages and time [1][3]. Efforts

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

*Corresponding author.

†These authors contributed equally.

miruna@ssn.edu.in (P. Mirunalini); sanjhay2110246@ssn.edu.in (V.Sanjhay); rohit2110090@ssn.edu.in (S. Rohitram); rohith2110565@ssn.edu.in (M.Rohith)

in creating a multi-lingual taxonomy further enhance computational understanding of sensory experiences [4].Furthermore, the development of a systematic theoretical framework to capture olfactory information from texts [5] represents a pivotal step toward automated systems for computational analysis.

3. Approach

For the Multimodal Understanding of Smells in Texts and Images (MUSTI) task[6][7][8], an integrated approach leveraging state-of-the-art language and vision models, specifically employing the Contrastive Language-Image Pretraining (CLIP) model [9] has been devised. The methodology involves the seamless fusion of textual and visual information to decipher and correlate olfactory references[10] within textual passages and corresponding images from historical periods ranging from the 16th to the 20th century. BERT model was also used for tokenization of the text provided in the dataset. BERT [11] is often used in natural language processing tasks like tokenization to break huge chunks of texts into meaningful units. Its tokenization process involves splitting text into sub-word units that capture intricate linguistic patterns for better contextual understanding during model training and inference.The backbone used for vision features extraction in CLIP models[9] like openai/clip-vit-base-patch16 is a Vision Transformer (ViT) architecture. This architecture represents images as sequences of patches and processes them through a transformer network, enabling the model to capture spatial relationships and features from the image.A classification threshold of 470 is used to determine whether the average similarity score between the image and text exceeds this threshold.

3.1. Data Preprocessing and Feature Extraction

The MUSTI dataset, a collection of copyright-free texts and images from various repositories and archives, forms the basis for this task. Annotated with over 80 categories of smell objects and gestures, the dataset allows participants to develop models aimed at recognizing and linking olfactory references across languages and modalities.

Initially, the dataset undergoes preprocessing of the provided textual data, comprising multilingual passages in English, German, Italian, and French, using a tokenization scheme implemented through the 'bert-base-multilingual-cased' tokenizer. To accommodate varying text lengths, the data is divided into chunk of texts that fit within the maximum sequence length compatible with the CLIP model's input requirements. Simultaneously, images linked to the textual descriptions are acquired by parsing the provided URLs. These images undergo resizing and conversion to numerical arrays for further processing.

To create feature representations conducive to model interpretation, a combination of pixel-level information from images and token embeddings derived from textual segments are utilized by our model. Image arrays are flattened, while textual segments are tokenized and flattened, yielding joint feature representations that capture the essence of both modalities.

3.2. Model Utilization: CLIP Integration for Multimodal Understanding

The proposed work employs the CLIP model which is designed to predict the image and text pairings. CLIP's embeddings for images and text share the same space and the CLIP's layers leverage a large-scale pretrained model which enables us to encode semantic relationships between text and images. This allows for cross-modal understanding, aiding in associating textual descriptions of smells with relevant visual representations, facilitating multimodal comprehension of olfactory concepts. The CLIP loss aims to maximize the cosine similarity

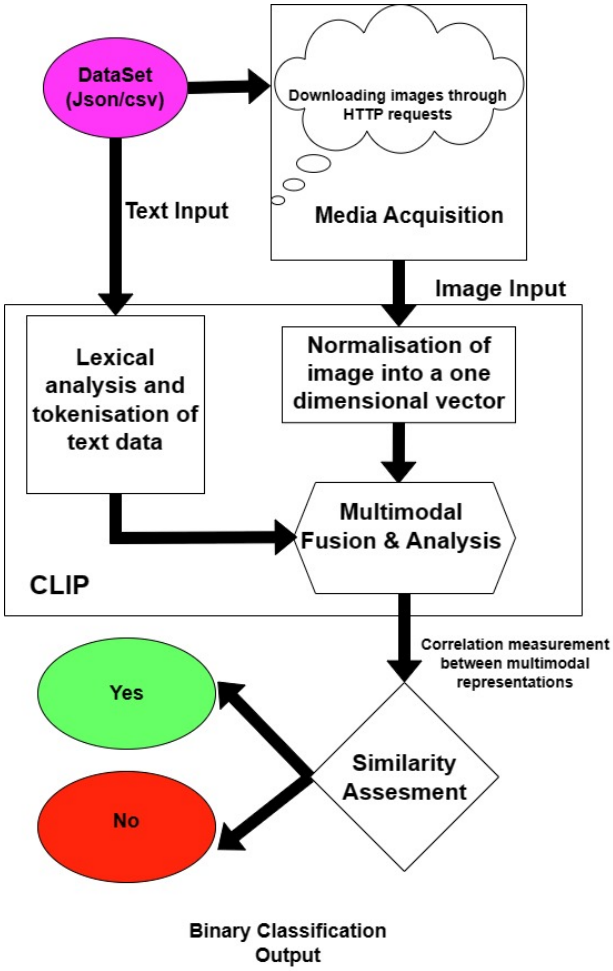

Figure 1: Architecture Diagram

between the image and text embeddings for the N genuine pairs in the batch while minimizing the cosine similarity for the N^2 - N incorrect pairings. This similarity measure serves as an indicator of the correspondence or shared olfactory source between the textual descriptions and the accompanying images.

3.3. Evaluation and Result Interpretation

The cosine similarities between text and images were computed and also an average similarity score across all segmented text chunks associated with each image is derived, providing a comprehensive assessment of the overall olfactory connection between text and image pairs. The average similarity scores are benchmarked against a threshold value to determine the presence or absence of a shared olfactory source between text and image pairs. By assigning binary labels based on the similarity threshold, the classification task (MUSTI Classification) for recognizing the co-relation of smells across modalities is facilitated.

4. Results and Analysis

A series of experiments was conducted, iterating over different stages of data preprocessing, feature extraction, and model utilization to achieve comprehensive olfactory understanding across text and image modalities. The proposed system was evaluated using precision, recall, and F1-score metrics and with overall accuracy. The results are depicted in the below table 1.The accuracy was found out to be 61.38%.

Class	Precision	Recall	F1-Score	Support
NO	0.8510	0.5313	0.6542	559
YES	0.4353	0.7953	0.5627	254
Accuracy		0.6138		
Macro Avg	0.6432	0.6633	0.6084	813
Weighted Avg	0.7211	0.6138	0.6256	813

Table 1
Metrics Evaluation with inclusion of title

On further attempts, exclusion of title in processing the similarity led to a better accuracy of 63.47%. This could imply confusion caused due to the inclusion of title. The results of the was depicted in the following table 2

Class	Precision	Recall	F1-Score	Support
NO	0.8075	0.6154	0.6985	559
YES	0.4444	0.6772	0.5367	254
Accuracy		0.6347		
Macro Avg	0.6260	0.6463	0.6176	813
Weighted Avg	0.6941	0.6347	0.6479	813

Table 2
Metrics Evaluation without inclusion of title

4.1. Interpretation of Results

The achieved accuracy of 63.47% in predicting the presence ('YES') or absence ('NO') of common smell sources between text passages and images demonstrates a moderate level of success in our model's performance. The precision, recall, and F1-scores for each class ('YES' and 'NO') indicate notable differences in the model's ability to predict positive and negative instances. The model displays a relatively higher precision of 80.75% for identifying cases where there are no common smell sources and a score of 61.54% which indicates that it misses a considerable number of actual instances where there are no common smell sources. The F1-score of 60.84% suggests a fair balance between precision and recall for the classes.

5. Discussion And Outlook

To refine the model, options include further fine-tuning CLIP and optimizing feature engineering techniques for better discernment of nuanced textual and visual olfactory relationships. Augmenting and enriching the dataset with diverse textual genres, images from various historical periods, and more smell-related annotations can significantly enhance the model's comprehension of olfactory references across different contexts and time frames.

References

[1] P. Lisena, D. Schwabe, M. van Erp, R. Troncy, W. Tullett, I. Leemans, L. Marx, S. C. Ehrich, Capturing the semantics of smell: The odeuropa data model for olfactory heritage information, in: European Semantic Web Conference, Springer, 2022, pp. 387–405.

[2] S. C. Ehrich, C. Verbeek, M. Zinnen, L. Marx, C. Bembibre, I. Leemans, Nose-first. towards an olfactory gaze for digital art history, in: CEUR Workshop Proceedings, volume 3064, CEUR Workshop Proceedings, 2021.

[3] S. Menini, T. Paccosi, S. Tonelli, M. Van Erp, I. Leemans, P. Lisena, R. Troncy, W. Tullett, A. Hürriyetoğlu, G. Dijkstra, et al., A multilingual benchmark to capture olfactory situations over time, in: Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, 2022, pp. 1–10.

[4] S. Menini, T. Paccosi, S. S. Tekiroglu, S. Tonelli, Building a multilingual taxonomy of olfactory terms with timestamps, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, 2022, pp. 4030–4039.

[5] S. Tonelli, S. Menini, Framenet-like annotation of olfactory information in texts, in: Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, 2021, pp. 11–20.

[6] H. Ali, T. Paccosi, S. Menini, Z. Mathias, L. Pasquale, A. Kiymet, T. Raphaël, M. van Erp, Musti-multimodal understanding of smells in texts and images at mediaeval 2022, in: Proceedings of MediaEval 2022 CEUR Workshop, 2022.

[7] A. Kiymet, H. Ali, T. Raphaël, T. Paccosi, S. Menini, Z. Mathias, C. Vincent, Multimodal and multilingual understanding of smells using vilbert and muniter, in: Proceedings of MediaEval 2022 CEUR Workshop, 2022.

[8] 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.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[10] B. Huber, T. Larsen, R. N. Spengler, N. Boivin, How to use modern science to reconstruct ancient scents, Nature Human Behaviour 6 (2022) 611–614.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).