SENSE-LM : A Synergy between a Language Model and Sensorimotor Representations for Auditory and Olfactory Information Extraction

Cédric Boscher
INSA Lyon
LIRIS – UMR 5205 CNRS
Lyon, France
cedric.boscher@insa-lyon.fr

Christine Largeron
Université Jean Monnet (UJM)
Laboratoire Hubert Curien — UMR 5516 CNRS
Saint-Etienne, France
christine.largeron@univ-st-etienne.fr

Véronique Eglin
INSA Lyon
LIRIS – UMR 5205 CNRS
Lyon, France
veronique.eglin@insa-lyon.fr

Elöd Egyed-Zsigmond
INSA Lyon
LIRIS – UMR 5205 CNRS
Lyon, France
elod.egyed-zsigmond@insa-lyon.fr

Abstract

The five human senses – vision, taste, smell, hearing, and touch – are key concepts that shape human perception of the world. The extraction of sensory references (i.e., expressions that evoke the presence of a sensory experience) in textual corpus is a challenge of high interest, with many applications in various areas. In this paper, we propose *SENSE-LM*, an information extraction system tailored for the discovery of sensory references in large collections of textual documents. Based on the novel idea of combining the strength of language model, BERT, and linguistic resources such as sensorimotor norms, it addresses the task of sensory information extraction at a coarse-grained (sentence binary classification) and fine-grained (sensory term extraction) level. Our evaluation of *SENSE-LM* for two sensory functions, Olfaction and Audition, and comparison with state-of-the-art methods emphasize a significant leap forward in automating these complex tasks.

1 Introduction

Sensoriality, as a psycho-physiological concept (Geldard, 1953), models the human perception of the world through the five Aristotelian sensory functions (Sorabji, 1971): *visual (VIS), gustatory (GUS), olfactory (OLF), auditory (AUD) and haptic (HAP).* A sixth sense, interoception (INT), was more recently introduced by Craig (2002), referring to the emotional and physical sensations inherent to the inside of the human body. Sensory linguistics refers to the studying of the relationship between human language and sensory experiences (Winter, 2019).

This research domain has many real-life applications, such as cognitive sciences, cultural history, or even urban planning. For instance, Murphy (2019) evidenced a strong relationship between the way olfactory experiences are expressed in the language of inpatients, and the chances of suffering from Alzheimer's disease. Pardoen (2019) focuses on the discovery of auditory indices in large document corpora to design a realistic reconstruction of the sound atmosphere of the City of Paris during the 19th century. Menini et al. (2022a) focuses on the sensory heritage of smells between the 17th and 20th century, with the goal of providing strong assets for museums to provide olfactory experiments for visitors. Such ambitious challenges may jointly solicit complementary spheres of competences, such as Art and Cultural History, Cognitive Sciences, and more recently, computational domains such as Semantic Web (Lisena et al., 2022) and Natural Language Processing (Mpouli et al., 2019; Menini et al., 2022b), with the interest of enhancing sensory information mining processes, notably with language models such as BERT (Devlin et al., 2019).

A set of lexical field generation approaches (Fast et al., 2016b; Tekiroglu et al., 2014; Mpouli et al., 2020) additionally provide interesting vocabulary resources referring to specific sensory domain, but employing them without integrating the text context may limit their scope to a very explicit level of sensory information. In parallel, a strong advance in the modeling of associations between concepts and sensory experiences has been opened by the appearance of the Lancaster Sensorimotor Norms (Lynott et al., 2020). This resource asso-

Findings of the Association for Computational Linguistics: EACL 2024, pages 1695–1711

ciates 40 000 English lemmas to the way they may evoke each sense, from a human judgement perspective. For instance, such a model represents the fact that, in essence, a concrete concept such as *"cat"* may evoke well-identified sounds and textures, and to a lesser extent odors, but probably no taste. Such resources provide strong assets on the sensory definition of concepts, but still lack of context-awareness, as they focus on isolated terms.

In this paper, we propose *SENSE-LM*, a novel system that combines the strengths of context-aware models such as language models (LM), linguistic resources, namely sensorimotor representations and lexical generation techniques, to provide a robust approach for detecting sensory-related information in large text corpora, at the sentence and term level.

We make the following contributions:

- We propose *SENSE-LM*, a sensory information extraction system working in two steps: Firstly, a coarse-grained binary classification step, that combines the strength of BERT and sensorimotor representations of words, to detect, within a textual corpus, sentences that explicitly evoke the presence of a given sensory function. Secondly, a fine-grained information extraction step, that extracts the precise terms referring to the evocation of the considered sensory function. The code and data are publicly available[1].
- Unlike existing works (Mpouli et al., 2019; Menini et al., 2022c), *SENSE-LM* is sensory-agnostic by design, i.e., it is not tailored for one specific sense. It may either be applied for the analysis of tastes, sounds, odors, or even textures, as its main components consider all senses.
- To evaluate the contributions of its different components, we conduct an ablative study of *SENSE-LM* for sensory information extraction, applied to two sensory functions, namely Olfaction and Audition. Moreover, a comparative evaluation of *SENSE-LM* with state-of-the-art solutions, and a bleeding-edge large language model, GPT-4 (OpenAI, 2023), confirms its good performances.
- To compensate the lack of benchmark datasets for this evaluation, we built an Auditory-oriented Artificial Dataset, generated with GPT-4 and manually labelled. We make publicly available a dataset of 1000 sentences with binary annotation (positive, i.e., containing a sound reference, or negative), including 500 positive sentences with a token-level annotation for terms expressing sound references.

The remainder of this paper is organized as follows. Section 2 provides an overview of the related work. Section 3 elaborates on the objectives and design principles of our contributions. We provide experimental evaluations and analysis in Section 4. We summarize our findings and draw our conclusions in Section 5, and discuss the current limitations of our solution in Section 6. An Appendix provides further analyses of our experiments.

2 Related Work

One of the main challenges of textual sensory information research, that we address in this paper, is about finding terms or expressions related to a sensory experience in a corpus of textual documents. In this section, we describe the existing approaches for addressing this task.

2.1 Lexical Resources Based Approaches

Lexical approaches intend to automatically build a list of terms or a taxonomy related to a specific sensory domain, from a small sample of seed terms. Lexifield (Mpouli et al., 2020), a system for automatic building of lexicons by semantic expansion of short word lists, was proposed and directly applied to the search for terms evoking either the auditory or olfactory sensory functions in literary works. This solution empirically dominates lexicon generation approaches such as Empath (Fast et al., 2016a) or Sensicon (Tekiroglu et al., 2014), by automatically enriching a small set of seed terms, with the help of techniques based on semantic similarity in embedding spaces (Bojanowski et al., 2017; Pennington et al., 2014) and external resources such as dictionaries in various target languages (Amsler, 1981; Sagot and Fišer, 2012). Such resources have been exploited for the automated detection of sound descriptions (Mpouli et al., 2019); the described approach happened to struggle with issues such as polysemy, but also provided encouraging, yet improvable results, as it considered including word embeddings at their premises, on the base of naive hypotheses.

[1] https://github.com/cfboscher/sense-lm

2.2 Language Models Based Approaches

Some preliminary works opened first contributions of sensory information mining based on language models. Menini et al. (2022b) solve a simple binary classification task corresponding to the following question: *"Considering a sentence s, does s contain a reference to olfaction ?"* with MacBERTh (Manjavacas and Fonteyn, 2021), a variant of BERT pre-trained on historical texts (1450–1950). Massri et al. (2022) propose a text mining method for detecting olfactory references and sentiments related to olfaction. They introduce a fine-grained olfactory concepts detection approach, but still based on naive hypotheses, as they use textual rules and only focus on objects and sentiments, which provides a potentially limited analysis of expressions of sensoriality.

As the efficiency of these solutions strongly depends on the quality of the ground truth labels and have a hardly explainable behavior (Zhao et al., 2023), they are difficult to exploit by non-specialists. They may require the support of domain specialists, both for annotating the data and for controlling the quality of results in a production environment.

Khalid and Srinivasan (2022) proposed a first approach based on a language model (BERT) to predict the most probable sensory function associated to a masked word in a sentence context. To generate its ground truth labels, this work involves the use of the Lancaster Sensorimotor Norms (Lynott et al., 2020), a linguistic resource of 40 000 English terms labelled according to their matching with each sensory function, but does not exploit them as classification features yet. Kennington (2021) first used sensorimotor norms as classification features, enriching a language model, ELECTRA (Clark et al., 2020), but for solving tasks that are not related to sensory information extraction.

2.3 Motivations for our Work

Considering the limits of the aforementioned existing techniques, our motivation for proposing *SENSE-LM* is to overstep the respective current blind spots of different sensory information approaches, and to bring a new step forward by combining the respective advantages of each family. Indeed, approaches based on language models provide an encouraging (yet perfectible) ability to embed a sentence context to detect the presence of a sensory function with a coarse-grained approach (Menini et al., 2022c), but it limits to contextual information, and does not include any linguistic resource describing sensoriality by design. It only considers that a concept may be sensory on the base of its context of utterance, without providing guarantees of understanding that a concept may evoke sensoriality in essence. In exchange, lexical resources (Tekiroglu et al., 2014; Fast et al., 2016b; Mpouli et al., 2020), and sensorimotor resources (Lynott et al., 2020) provide extensive knowledge of terms that may explicitly or implicitly be related to the presence of a given sensory function. These are interesting resources for fine-grained sensory reference detection, but their main weakness is that they still lack of context awareness and may struggle with challenging issues such as polysemy (Ravin and Leacock, 2000; Falkum and Vicente, 2015). More generally, labelling sensory references manually is a time-consuming task, that may even require multidisciplinary expertise, as suggested by Menini et al. (2022a). In this paper, we introduce *SENSE-LM*, a system that automatically extracts sensory references from text, by exploiting the complementary advantages of language models and lexical resources-based approaches. We experimentally show that they can work in synergy to overcome the current limits of sensory information extraction techniques.

3 Methodology

In this section, we present our system *SENSE-LM*, designed for detecting text information describing sensory experiences in documents. *SENSE-LM* allows extracting sensory references in large corpora, at a sentence and at a token level. Figure 1 depicts its global workflow. Step 1 performs a coarse-grained classification task, aiming at identifying, within a set of documents D, the sentences that evoke the presence (or not) of references related to one of the five sensory functions. We may sum up this binary classification task by the following question: *"Does a sentence **s** expresses an idea evoking a given sense **m** among the five senses ?"* Then, Step 2 applies a fine-grained classification process, for extracting word utterances that reflect the presence of the target sensory function in the sentence context. We sum up this classification task with the following question : *"Which words, in this sentence **s**, evoke the presence of the given sensory function **m** ?"*.

It is worth noting that such a method adresses

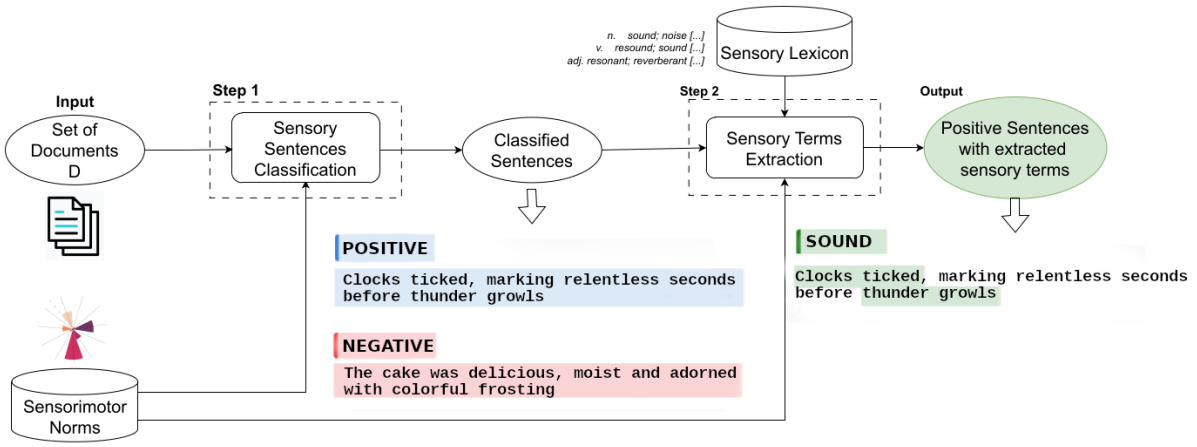

Figure 1: Global Workflow of *SENSE-LM* with an example for the sensory function Audition

the task of researching multi-sensory information, i.e. finding, in a same document, information that refer to several sensory functions. In that case, it is enough to apply a One-vs-Rest strategy, which consists to split the multi-class classification problem into several binary classification problems, one per class, and to learn a model on each. Thus, for instance, if a sentence contained several sensory information, it will classified positively by several instances of the model, whereas if it does not contain any, it will be classified negatively by all the models.

3.1 Step 1 — Sensory Sentence Classification

In the following, we describe our binary sentence classification model, considering text features extracted by BERT and a sensorimotor representation, implementing 11 human judgement based continuous values that we describe below:

Definition of the Sentence Classification Problem. We consider the ensemble of sensory functions $\mathbb{M} = \{\text{OLF, GUS, AUD, VIS, HAP, INT}\}$, corresponding to Olfactory, Gustatory, Auditory, Visual, Haptic and Interoceptive. We define a corpus D of textual documents composed of sentences. For each sensory function m of $\mathbb{M}$, each sentence $s \in D$ has a class label $C(s)$ which is positive (1) if it contains explicit references to m; otherwise its class label is negative (0). For instance, if we consider $m =$ AUD, *"Clocks ticked, marking relentless seconds before thunder growls."* is a positive sentence whereas *"The cake was delicious, moist, and adorned with colorful frosting"* is negative. This first step of *SENSE-LM* consists in classifying correctly the sentences according to the chosen sensory function m; which amounts to learning a classification function ϵ that maps each sentence s to a class label: $\epsilon : D \rightarrow \{1, 0\}$ s.t. $\epsilon(s) = C(s), \forall s \in D$.

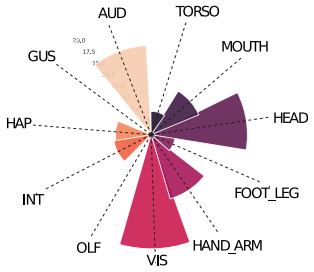

Figure 2: Sensorimotor representation of the sentence *"Clocks ticked, marking relentless seconds before thunder growls"*, plotting the sensory and motor functions.

Sensorimotor Representation Function. As a premise to the description of our solution, we present the concept of Sensorimotor Representation, based on the Lancaster Sensorimotor Norms (Lynott et al., 2020). This resource consists of an extensive set of 40 000 English lemmas evaluated by human annotators, asked to rate from 0 to 5 the semantic matching of a given lemma with 6 human sensory functions (the five Aristotelian Senses and the Interoception), and 5 motor functions corresponding to the usage of body parts (Mouth, Head, Torso, Arms / Hands, Legs / Feet). In other words, each lemma can be represented into a sensorimotor representation, i.e., an 11-dimensional vector of real values between 0 and 5, with 6 dimensions corresponding to the sensory functions, and 5 to the motor functions. Algorithm 1 details the calculation method to obtain the sensory representation of a sentence s depicted in Figure 2. We denote by L_{SN} a dictionary corresponding to words available in the Lancaster Sensorimotor Norms:

it maps each word w in s with its sensorimotor representation w_{SN} as an 11-dimensional vector $w_{SN} = (w_{SN}(j), j = 1, \ldots, 11)$, where $w_{SN}(m)$ corresponds to the component of w_{SN} associated with the sensory function $m \in \mathbb{M}$. The sensorimotor representation w_{SN} of w equals $lemma(w)_{SN}$ if the lemma associated to w exists in L_{SN}. In case this lemma is not included in L_{SN}, we consider the first element belonging to the set $Synsets(w)$ of WordNet synsets of w as defined by Miller (1995), i.e., synonymous words. Finally, if there is also no synset of w included in L_{SN}, the sensorimotor representation of w is an 11 dimensional vector with null components. As detailed in the algorithm, having determined this sensorimotor representation for each word $w \in s$, the sentence sensorimotor representation $s_{SN} = (s_{SN}(j), j = 1, \ldots, 11)$ of s is obtained by summing these word vectors.

Algorithm 1 Sensorimotor Representation

Input: Sentence s, Sensorimotor Norms L_{SN}
Output: Sensorimotor representation s_{SN}

$s_{SN} \leftarrow (0, 0 \ldots 0)$
$s \leftarrow RemoveStopWords(s)$
for $w \in s$ **do**
  if $lemma(w) \in L_{SN}$ **then**
    $v \leftarrow lemma(w)_{SN}$
  else
    $v \leftarrow (0, 0 \ldots 0)$
    for $i \in$ Synsets(w) **do**
      if $lemma(i) \in L_{SN}$ **then**
        $v \leftarrow lemma(i)_{SN}$
        $break$
      end if
    end for
    $s_{SN} \leftarrow s_{SN} + v$
  end if
end for
  return s_{SN}

Description of the Sentence Classification Model. The first step of *SENSE-LM* combines the sentence context awareness of BERT, and the knowledge on sensoriality provided by the Lancaster Sensorimotor Norms (Lynott et al., 2020). The latter has been proven to be a robust representation, providing a singular level of semantic similarity between terms, complementary to state-of-the-art embeddings (Wingfield and Connell, 2022). As shown in Figure 3, *SENSE-LM* takes a sentence s as input.

Its first branch implements BERT's successive stages: Embedding, Transformers and Pooler layers, which extracts an embedded representation of s of size 768 , denoted s_B.

The second branch of the model transforms the sentence s into its sensorimotor representation s_{SN},

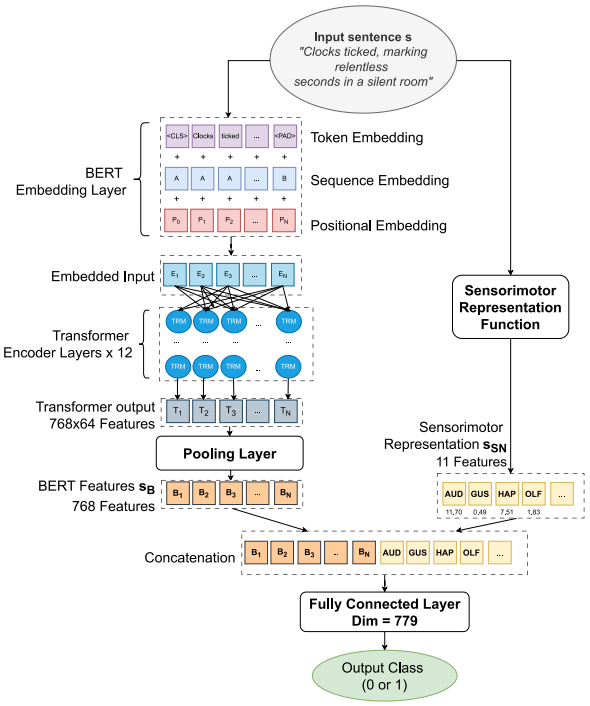

Figure 3: Model architecture for Step 1 of *SENSE-LM*

following the procedure detailed in Algorithm 1, which results in a vector of size 11.

Finally, the model concatenates s_B and s_{SN} into a global representation, and feeds it into a Fully-Connected layer (dimension = 779) that outputs either 1 if s is considered as sensory w.r.t. the sensory function m, or 0 if not.

3.2 Step 2 — Sensory Terms Extraction

Definition of the Sensory Terms Extraction Problem. The objective of the second step of *SENSE-LM* consists in extracting the tokens that refer to the expression of a given sense $m \in \mathbb{M}$ in a sentence s, within sentences classified positively in Step 1. We consider the following types of sensory terms, defined by the categories proposed by Menini et al. (2022a):

- Sensory word – Words that explicitly describe the presence of the target sensoriality: *"What was this **sound** ? [. . .]"*
- Sensory Source – Entities that create the sensoriality: *The cry of **a baby** [. . .]*
- Quality: "What a **horrible** smell [. . .]"
- Evoked Experience: "The taste of this cake **gave me nausea** [. . .]"

For each sensory function m of $\mathbb{M}$ and each sentence s belonging to D_{pos}, the set of sentences classified positively during the previous step, s is split into a sequence of tokens, denoted $t(s)$. To ensure

that the length of t remains constant for all positive sentences, we apply a padding, i.e., we fix a length l that corresponds to the length of the longest positive sentence, and in case $len(t(s)) < l$, we append k padding tokens denoted as <PAD> at the end of $t(s)$, with $k = l - len(t(s))$.

Each token $i \in t(s)$, excluding the padding tokens, has a ground truth class label $F(i, m)$ which is positive (1) if the token i refers to the sensory function m in the context of the sentence, and negative instead (0). We aim to learn a token classification function γ that takes $t(s)$ as an input, and returns a vector of class labels (1 or 0) for each $i \in t(s)$ such that:

$\gamma : t(s) \rightarrow (\{1, 0\}, \forall i \in t(s))$ s.t. $\gamma(t(s), m) = (F(i, m), \forall i \in t(s)), \forall s \in D_{pos}$

For instance, if we consider the sensory function $m = AUD$ and the sentence $s =$ *"Clocks ticked, marking relentless seconds before thunder growls."*, we obtain $t(s) =$ (**Clocks, ticked**, marking, relentless, seconds, before, **thunder, growls**, ..., <PAD>), where terms in bold reflect the presence of the sensory function m i.e. positive terms.

Our objective is then to learn the function γ which gives for this example:

$\gamma(t(s), m)$ = (**1, 1**, 0, 0, 0, 0, **1, 1**, ..., <PAD>)

Description of the Sensory Term Extraction Model. To address this task, we introduce a combinatorial approach involving three complementary steps :

Step 2.1. Term Classification with RoBERTa. Firstly, we propose to fine-tune a language model on the task of extracting sub-phrases in sentences that express the presence of a given sensoriality, by following the intuition of Dash (2021) who formerly addressed the task of identifying the terms that best reflect the main sentiment (Positive, Neutral, or Negative) expressed by tweets[2]. By analogy, we use a similar principle to detect words that best reflect the presence of the target sensoriality m.

We use a BERT architecture, with the RoBERTa pre-trained parameters set (Liu et al., 2019), that empirically shows improved performances on the task of classifying sensory and non-sensory tokens within a sentence context.

[2] https://www.kaggle.com/competitions/tweet-sentiment-extraction/leaderboard

Our input is the tokenized sentence t(s), and the predicted output is a vector denoted $V(t(s), m)$, with ones for positively predicted terms corresponding to the sensory function m, and zeroes for negatives. Thus, this first stage allows extracting a first set of words, classified as positive in the context by RoBERTa. $P_{pos}(s, m)$ denotes the set of words in $t(s)$ that map the words classified positively in $V(t(s), m)$, and $P_{neg}(s, m)$ the negative ones.

Step 2.2. Expansion with Lexical Resources. Secondly, we use a lexical resource, such as Lexifield (Mpouli et al., 2020) with the goal of expanding the list of sensory tokens preliminarily extracted in step 2.1. This lexicon denoted $\mathbb{L}_m$ contains a set of words belonging to the lexical field of the target sensory function m. For instance, we may consider $\mathbb{L}_{\text{OLF}} =$ {odour (noun) , smell (verb), ...} if $m =$ OLF.

For each word $w \in P_{neg}(s, m)$, we switch the corresponding value in $V(t(s), m)$ to 1 if $w \in \mathbb{L}_m$.

Step 2.3. Language and Human Judgement-Based Heuristic. Finally, with the objective of recovering false negative words omitted by the first classification step, and at the same time, avoiding introducing false positive examples significantly, we settle a heuristic that both considers the sensorimotor representation of candidate terms and their semantic proximity with positive examples. We denote by $\mathbb{E}$ a set of semantic embedding spaces, and $\text{CosSim}_e(a, b)$ the cosine similarity measure between words a and b in an embedding space $e \in \mathbb{E}$. For each word $w \in P_{neg}(s, m)$, we switch the corresponding value in $V(t(s), m)$ to 1 in case it combines the two following conditions:

1. $w_{SN}(m) > T$
2. $\exists e \in \mathbb{E}$, and $\exists x \in P_{pos}(s, m)$, s.t. $\text{CosSim}_e(w, x) > U$

where $w_{SN}(m)$ denotes, in the sensorimotor representation of the word w, the dimension corresponding to the sensory function m.

Condition 1 first ensures that the candidate term is coherent with the target sensory function m in essence. T defines the minimal threshold value of $w_{SN}(m)$, with $T \in [0, 5]$. Then, Condition 2 ensures that classifying w as positive makes sense in context, as it is semantically close to at least one

of the positive terms. $U \in [0, 1]$ defines the minimal cosine similarity value between a candidate term and at least one of the positive terms. Both T and U are tuned manually on the base of empirical analyses, although they could be determined by a grid search. At the end of this stage, the system returns the output $\gamma(t(s), m) = V(t(s), m)$.

4 Experiments and Analyses

This section presents an experimental evaluation of the effectiveness of *SENSE-LM*. The performances are measured for each step and compared with those provided by baselines that address the same task. An ablative study is also carried out to evaluate the interest of each of the components implemented in Step 2. The software and hardware environments of these experiments are described in Appendix B, and an analysis of the computational costs of *SENSE-LM* is provided in Appendix D.

4.1 Datasets

Our experiments are performed on two datasets: **Odeuropa: English Benchmark**[3] **(Menini et al., 2022c)** This state-of-the-art dataset focused on olfactory experiences from the 17th to the 20th century. It contains 2176 sentences with a positive sentence ratio of 0.28 and, 5530 utterances of smell related terms, distributed in 602 sentences.

Auditory-oriented Artificial Dataset. Due to the lack of sensory dataset corresponding to other sensory functions and including consistent annotation, we built an artificial dataset composed of synthetic sentences generated with GPT-4 (OpenAI, 2023) and containing references to sounds. We carefully ask GPT-4 to create examples respecting a realistic diversity of sentence structures with different sentence lengths (400 sentences of maximum 10 words, 400 sentences of between 25 and 35 words, and 200 sentences between 35 and 50 words) with a ratio of positive sentences examples of 0.5. Our generation protocol is detailed in Appendix F.1.

Then, the sensory terms appearing in positive sentences (500 sentences) have been labelled using Label Studio (Tkachenko et al., 2020-2022) by a European PhD student, with the following instruction : "Label terms that either evoke the production of sounds, sound producers entities, qualities related to sound experiences or evoked sound experiences", followed by the examples provided in Section 3.2. The dataset is publicly available[4].

4.2 Experimental Settings

The datasets have been split into training and test sets, with a ratio of 0.2 for the test set. Our models and the baselines are trained on the same data, with a 10-fold cross validation, and 5 experiment runs. The train / test splits and cross validation folds are generated using the same random seed value fixed to 42. We use the AdamW optimizer (Loshchilov and Hutter, 2017), with hyperparameters $lr = 2e^{-5}$ and $\epsilon = 1e^{-8}$, determined experimentally. The models are trained over 30 epochs. The evaluation measures are the Macro Precision, Recall and F1-Score, and the reported results correspond to the average scores, with standard deviation, computed over all runs.

4.3 Evaluation of Step 1 — Binary Sentence Classification

First, we evaluate the performances of the binary classifier implemented in Step 1 of *SENSE-LM* for detecting correctly the presence or not of a sensory function m at the sentence level.

Model Setting The BERT component of our architecture considers, for each dataset, respective pre-trained parameters, determined on the base of empirical observations : for the Odeuropa dataset (historical texts), as recommended by (Menini et al., 2022c), we use MacBERTh's pre-trained parameters that provide the best results. For the Auditory dataset (contemporary texts), we use the default `bert-base-uncased`[5] parameters.

Baselines First, we compare *SENSE-LM* with a simple BERT model with the same pre-trained parameters as the ones provided to the BERT component of our architecture. Then, we compare with a scenario in which sentences are only described with the sensorimotor representation (11 features), and classified by a Logistic Regression. We denote this second baseline by LR(s_{SN}). As GPT-4 allegedly comes with high potential for handling a large panel of NLP tasks, we also compare the efficiency of our solution against such a model for this classification task. We ask GPT-4 to solve this sensory sentence classification task, by first showing it examples, corresponding to the training set,

[3] https://github.com/Odeuropa/benchmarks_and_corpora.

[4] https://github.com/cfboscher/sense-lm

[5] https://huggingface.co/bert-base-uncased

and asking it to classify unseen examples, corresponding to our test set. The protocol implemented with GPT-4 is detailed in Appendix F.2.

Results The results presented in Table 1 show that *SENSE-LM* obtains better performances for both datasets, compared to the concurrent baselines (BERT classifier, LR(s_{SN}) and GPT-4), for the Precision, Recall and F1-Score measures.

In the case of the Odeuropa dataset, we notice close performances between BERT and GPT-4; the latter offers a precision equivalent to BERT, and a recall marginally below. Such a behaviour may result from the tangible limits of the information level that language models such as BERT or GPT-4 can infer from text, missing the inclusion of a human judgement based projection of concepts, contrary to the guarantees offered by *SENSE-LM*. Moreover, as the dataset is relatively small (2176 sentences), containing heterogeneous sources of documents from different eras, generalizing the classification problem on the base of the vocabulary only may be a difficult task, even for a large language model such as GPT-4. Then, OpenAI (2023) do not delve into details about the pre-training data of GPT-4, and do not provide guarantees on its real ability to work with historical data such as the Odeuropa dataset, which is a reasonable explanation on why GPT-4 may not work as well as MacBERTh, and *SENSE-LM* by extension.

In exchange, *SENSE-LM* reaches a F1-Score of 93.16% for Odeuropa and 97.12% on the Auditory dataset, dominating the compared baselines. This confirms the interest of enriching the model's training by integrating the sensorimotor representation to its architecture, for detecting the presence of a sensory function within a sentence.

4.4 Evaluation of Step 2 — Sensory Terms Extraction

This second set of experiments aims to evaluate the effectiveness of the term extraction from the sentences classified positively in the previous step.

Model Setting According to the sub-steps described in Section 3.2, we set our model as follows:

In Step 2.1, we set up the BERT component with the RoBERTa pre-trained parameters, and we fine-tune the model on our dataset. The fine-tuned model is used for predicting a first set of words for each candidate sentence.

In Step 2.2, as a lexical resource, we consider the lexicons of sensory words provided by Lexifield (Mpouli et al., 2020). In the case of Olfaction, the lexicon contains 155 English terms explicitly evoking smell experiences, and for Audition, 551 words evoking auditory experiences, including common names, verbs, and adjectives.

In Step 2.3, we configure our heuristic by including three embeddings in our set $\mathbb{E}$: `'word2vec-google-news-300'` (Church, 2017), `'glove-wiki-gigaword-300'` (Sakketou and Ampazis, 2020), and the sensorimotor representation defined in Section 3.1. We set the threshold values $T = 3.50$ and $U = 0.65$ for the Odeuropa dataset and, $T = 4.50$ and $U = 0.75$ for the Auditory dataset, which empirically correspond to optimal values estimated through a series of experiments.

Baselines We compare the performances of the second step of *SENSE-LM* with a simple lexicon-based baseline, denoted Lexifield($\mathbb{L}_{\text{m}}$); we consider a naive scenario in which all term utterances that are included in $\mathbb{L}_{\text{m}}$ are labelled positive, and the others are labelled negative. We also compare *SENSE-LM* with a stand-alone RoBERTa classifier and with GPT-4 using the same principle as in Section 4.3. A detailed description of our protocol is available in Appendix F.3.

Results Table 2 presents the results provided by the baselines (on top) and by *SENSE-LM*, with an ablative evaluation of each component (on bottom). *SENSE-LM* shows the best overall performances. For the Odeuropa dataset, *SENSE-LM* outperforms the F1-Score of Lexifield by 22% and the F1-score of RoBERTa alone by more than 5%. *SENSE-LM* also improves by 2% the F1-score of RoBERTa for the Auditory Dataset. The gap between RoBERTa and *SENSE-LM* is lower in this case; as the Auditory dataset contains synthetic data, it may include sentence construction patterns, which make the term extraction task easier even for the RoBERTa classifier alone, reducing the added value of our architecture, although it remains visible.

GPT-4 performs better than Lexifield, but still struggles with this task, with a F1-Score barely over 60%. Our reasoning on the performance limits of GPT-4 detailed in Section 4.3 may remain valid in this new case, and even be accentuated by the even smaller data sample used for the training task, as we only dispose of 600 sentences, using only 80% of them for the training. In such conditions, and without any guarantee on the abilities of GPT-4 to distinguish olfactory concepts from a hu-

	Odeuropa Benchmark Dataset			Auditory Artificial Dataset		
Method	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
BERT	91.51 ± 1.12	90.12 ± 0.61	90.80 ± 0.85	96.03 ± 0.31	96.14 ± 0.64	96.08 ± 0.45
LR(s_{SN})	82.25 ± 1.51	72.33 ± 1.22	76.97 ± 1.36	87.64 ± 1.14	87.04 ± 1.32	87.23 ± 1.23
GPT-4	91.59 ± 1.04	89.42 ± 2.21	90.4 ± 1.61	N/A*	N/A*	N/A*
SENSE-LM	**94.09 ± 0.81**	**92.26 ± 0.72**	**93.16 ± 0.76**	**97.01 ± 0.15**	**97.22 ± 0.24**	**97.12 ± 0.19**

Table 1: Evaluation of *SENSE-LM*'s binary sentence classification step versus baselines.

	Odeuropa Benchmark Dataset			Auditory Artificial Dataset		
Method	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Lexifield ($\mathbb{L}_m$)	77.3 ± 1.33	43.53 ± 1.17	55.69 ± 1.25	43.25 ± 0.18	16.32 ± 0.27	23.69
GPT-4	52.90 ± 2.11	70.99 ± 2.36	60.62 ± 2.24	N/A*	N/A*	N/A*
SENSE-LM (Step 2.1)	80.01 ± 2.22	66.32 ± 1.13	72.52 ± 1.68	91.51 ± 2.84	89.25 ± 2.94	90.36 ± 2.89
SENSE-LM (Step 2.1 ∪ Step 2.2)	81.5 ± 2.11	72.7 ± 1.56	76.84 ± 1.74	**91.75 ± 2.84**	92.49 ± 2.75	92.11 ± 2.81
SENSE-LM (Step 2.1 ∪ Step 2.3)	80.48 ± 1.65	70.21 ± 1.87	74.99 ± 1.77	91.19 ± 2.76	92.32 ± 2.81	91.75 ± 2.79
***SENSE-LM* (All steps)**	**82.01 ± 1.81**	**73.62 ± 1.56**	**77.58 ± 1.65**	91.65 ± 2.72	**93.01 ± 2.65**	**92.32 ± 2.70**

Table 2: Evaluation of *SENSE-LM*'s sensory terms extraction step versus baselines.

* As the Auditory Dataset was generated using GPT-4 itself, on the base of an explicit definition of our classification criterion, we do not consider evaluating the classification of the latter model on this dataset, as it would provide biased results.

man judgement perspective and to handle properly historical texts, we may have a reasonable explanation on why GPT-4 does not work well on this task. Indeed, our additional experiments in Appendix E show the importance of benefiting from sensorimotor representations in order to detect sensoriality, particularly when working with a small training dataset. A reasonable explanation for the high improvement brought by *SENSE-LM* is that we additionally require the sentence context and a human judgement-based representation of concepts to better identify the relationship between an explicit odor, and contextually related entities.

In a second time, the ablative evaluation of *SENSE-LM* highlights the interest of combining successively its 3 steps, as including all of them in a unique framework provides the highest results.

Appendix C provides an error analysis of *SENSE-LM*, detailing its performances scores grouped by part-of-speech and by semantic category (as defined in Section 3.2), in order to highlight its strengths and weaknesses.

5 Conclusion and Future Works

In this paper, we presented *SENSE-LM*, a novel framework for coarse-grained, at the sentence level, and fine-grained, at the word level, sensory references detection. As far as we know, *SENSE-LM* is the first approach proposing a combination of sensorimotor representations with the text features of language models such as BERT for sensory information extraction in text documents. In addition, unlike other systems which are dedicated to a particular sensoriality, it offers the advantage of being generic and applicable to any sense.

Its evaluation on two datasets for two different sensory functions, Olfaction and Audition, provides enhanced and encouraging results compared to state-of-the-art solutions. Moreover, an ablative study confirms the contribution of each component of the system, highlighting that using sentence context-aware approaches and human-judgement based approaches together brings a new step forward in the task of identifying sensory references in text, as these two approaches are complementary.

This work opens interesting directions for future works. Our approach, evaluated on a sensory information research task, could be transferred to similar tasks involving human judgement, such as sentiment analysis or political polarity analysis, by replacing the sensorimotor representation function by an equivalent function built on human-judgement based resources tailored for other domain-specific tasks. Thus, our work on sensoriality shows a new way to enhance a human judgement oriented task with the help of multimodality, and opens a set of interesting research directions for other application domains. From a language-models study perspective, we may inspire from existing works that enrich language models with extra modalities such as images alongside sensorimotor representations (Kennington, 2021). The principle of combining the three aforementioned modalities (text, image and sensorimotor), has been applied to purely text-oriented tasks, but has not been applied yet to the research of sensory indices in text corpora. Conversely, the synergy of text and sensorimotor modalities, that we valued in this paper, could be employed to enrich computer vision and multi-modal architectures for extracting visual sensory information from images.

6 Limitations

Although it shows promising results, the usage of *SENSE-LM* may suffer from operational limitations, either related to its design or to its adaptation to use-cases. Firstly, the strength of *SENSE-LM* against existing approaches resides, to an important extent, in the integration of Sensorimotor Norms; the latter resource provides interesting added value in the accomplishment of our tasks, but it is worth noting that on the day of writing, Sensorimotor Norms exist for a limited vocabulary, namely, lemmas known by 80% of a group of subjects representative of the English-speaking community (Lynott et al., 2020). It covers a wide spectrum of current vocabulary, but such a resource may become hard to exploit for rare and domain-specific vocabulary.

Yet, the research of sensory references may be solicited for specialized scientific research areas such as chemistry (Brate et al., 2020), that involve uncommon and domain-specific vocabulary that may have no equivalent synset in WordNet (Miller, 1995). For instance, in the chemistry area, the term *chalcogen* designates a family of metals that may evoke specific smells, such as sulfur (Vogel et al., 2019). Notwithstanding, the word *chalcogen* is neither listed in Sensorimotor Norms, nor on WordNet, which makes it a blind spot in the scope of *SENSE-LM* by default. An alternative solution would be to include domain-specific terms in the lexical resource component, but it supposes a prior exhaustive definition of terms related to the application domain, or even the usage of knowledge bases.We may face a similar issue for analyzing historic texts. Indeed, *SENSE-LM*'s Sensorimotor Representation function only covers 87% of unique terms appearing in the Odeuropa dataset (which corresponds to 94% of word utterances in the whole corpus), while replacing values for missing words by zeroes. This coverage may decrease in case we apply our system to even older texts (before the 17th century).

Additionally, Sensorimotor Norms are predominantly available for the study of the English language. Preliminary works have been provided for Dutch (Speed and Brybaert, 2021), Chinese (Zhong et al., 2022) and French (Lakhzoum et al., 2023), but for instance, the latter only covers 1,100 words, while the French language counts over 38 000 words (Ferrand et al., 2010). This makes *SENSE-LM*, to some extent, suitable for English but hardly adaptable to other languages by design, until consequent sensorimotor resources are released.

At the time of writing, it is difficult to benchmark to what extent the effectiveness of *SENSE-LM* is generalizable. Even if our system may be useful in many use cases in practice, evaluating our solution on real data is difficult, as far as consequent and labelled datasets are too few in numbers until now; only the Odeuropa benchmark dataset (Menini et al., 2022c), as a public dataset coming with a ground truth annotation, suits our needs for experimenting our solution. Thus, our experimentation on real data has been practically limited to one sensory function in this paper, olfaction, although it has also been evaluated on artificial data for another sensory function, confirming its ability to deal with different functions The construction of suitable datasets may be considered for several applications, but labelling correctly sensory references is a hard task, as it requires a high human effort and involves in-depth knowledge of the application domains. The release of datasets providing sensory information dedicated to the other sensory functions would be a strong asset to push our method a step farther, for example by considering multisensory classification at a sentence and a token level. Constructing a valuable ground truth is still a difficult task, as transdisciplinary projects such as Odeuropa (Menini et al., 2022a) or Polifonia[6] require the intervention of domain experts in several research areas such as history, musicology or cognitive sciences. In Appendix E, we discuss the performances of our system depending on the size of the available training data.

Ethics Statement

All datasets and code used in this work are released publicly under open-source licenses, and do not contain any personal information.

Our system aims to reproduce the classification of human annotators, on the base of a few examples. Thus, biases may be reproduced by our models. Furthermore, as we work with historical data, it may contain outdated and controverted expressions that do not reflect the authors' opinion.

At the same time, as we work with artificial data generated by GPT-4, the synthetic data we use in our study may express objectively erroneous facts, as GPT-4 does not integrate any notion of fact-checking regarding generated contents.

[6] https://polifonia-project.eu/

Acknowledgements

We warmly acknowledge the french Auvergne-Rhône-Alpes Region for their support of the *Symtesens* project under the *Pack Ambition Research 2020-2024 Program*. This project involves the collaboration of the french National Scientific Research Center CNRS, three academic research teams and the french heritage institution, *Lyon Municipal Archives*.

References

Robert A Amsler. 1981. A taxonomy for english nouns and verbs. In 19th Annual Meeting of the Association for Computational Linguistics, pages 133–138.

Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5:135–146.

Ryan Brate, Paul Groth, and Marieke van Erp. 2020. Towards Olfactory Information Extraction from Text: A Case Study on Detecting Smell Experiences in Novels. ArXiv:2011.08903 [cs].

Kenneth Ward Church. 2017. Word2vec. Natural Language Engineering, 23(1):155–162.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

Arthur D Craig. 2002. How do you feel? interoception: the sense of the physiological condition of the body. Nature reviews neuroscience, 3(8):655–666.

Dash. 2021. Extract the right Phrase From Sentence. Medium, Analytics Vidhya.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [cs].

Ingrid Falkum and Agustin Vicente. 2015. Polysemy: Current perspectives and approaches. Lingua, 157.

Ethan Fast, Binbin Chen, and Michael Bernstein. 2016a. Empath: Understanding Topic Signals in Large-Scale Text. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pages 4647–4657. ArXiv:1602.06979 [cs].

Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016b. Empath: Understanding topic signals in large-scale text. In Proceedings of the 2016 CHI conference on human factors in computing systems, pages 4647–4657.

Ludovic Ferrand, Boris New, Marc Brysbaert, Emmanuel Keuleers, Patrick Bonin, Alain Méot, Maria Augustinova, and Christophe Pallier. 2010. The french lexicon project: Lexical decision data for 38,840 french words and 38,840 pseudowords. Behavior research methods, 42:488–496.

Frank A Geldard. 1953. The human senses. Wiley.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Casey Kennington. 2021. Enriching Language Models with Visually-grounded Word Vectors and the Lancaster Sensorimotor Norms. In Proceedings of the 25th Conference on Computational Natural Language Learning, pages 148–157, Online. Association for Computational Linguistics.

Osama Khalid and Padmini Srinivasan. 2022. Smells like Teen Spirit: An Exploration of Sensorial Style in Literary Genres. ArXiv:2209.12352 [cs].

Dounia Lakhzoum, Marie Izaute, and Ludovic Ferrand. 2023. Word-association norms for 1,100 french words with varying levels of concreteness. Quarterly Journal of Experimental Psychology, page 17470218231154454.

Pasquale Lisena, Daniel Schwabe, Marieke van Erp, Raphaël Troncy, William Tullett, Inger Leemans, Lizzie Marx, and Sofia Colette Ehrich. 2022. Capturing the Semantics of Smell: The Odeuropa Data Model for Olfactory Heritage Information. In Paul Groth, Maria-Esther Vidal, Fabian Suchanek, Pedro Szekley, Pavan Kapanipathi, Catia Pesquita, Hala Skaf-Molli, and Minna Tamper, editors, The Semantic Web, volume 13261, pages 387–405. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. ArXiv, abs/1711.05101.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. Behavior Research Methods, 52(3):1271–1291.

Enrique Manjavacas and Lauren Fonteyn. 2021. Macberth: Development and evaluation of a historically pre-trained language model for english (1450-1950). pages 23–36.

M. Besher Massri, Inna Novalija, Dunja Mladenić, Janez Brank, Sara Graça da Silva, Natasza Marrouch, Carla Murteira, Ali Hürriyetoğlu, and Beno Šircelj. 2022. Harvesting Context and Mining Emotions Related to Olfactory Cultural Heritage. Multimodal Technologies and Interaction, 6(7):57.

Stefano Menini, Teresa Paccosi, Serra Sinem Tekiroglu, and Sara Tonelli. 2022a. Building a multilingual taxonomy of olfactory terms with timestamps. Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 4030–4039.

Stefano Menini, Teresa Paccosi, Sara Tonelli, Marieke Van Erp, Inger Leemans, Pasquale Lisena, Raphael Troncy, William Tullett, Ali Hürriyetoğlu, Ger Dijkstra, Femke Gordijn, Elias Jürgens, Josephine Koopman, Aron Ouwerkerk, Sanne Steen, Inna Novalija, Janez Brank, Dunja Mladenic, and Anja Zidar. 2022b. A multilingual benchmark to capture olfactory situations over time. In Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, pages 1–10, Dublin, Ireland. Association for Computational Linguistics.

Stefano Menini, Teresa Paccosi, Sara Tonelli, Marieke Van Erp, Inger Leemans, Pasquale Lisena, Raphael Troncy, William Tullett, Ali Hürriyetoğlu, Ger Dijkstra, Femke Gordijn, Elias Jürgens, Josephine Koopman, Aron Ouwerkerk, Sanne Steen, Inna Novalija, Janez Brank, Dunja Mladenic, and Anja Zidar. 2022c. A Multilingual Benchmark to Capture Olfactory Situations over Time. In Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, pages 1–10, Dublin, Ireland. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: A lexical database for english. Commun. ACM, 38(11):39–41.

Suzanne Mpouli, Michel Beigbeder, and Christine Largeron. 2020. Lexifield: a system for the automatic building of lexicons by semantic expansion of short word lists. Knowledge and Information Systems, 62(8):3181–3201.

Suzanne Mpouli, Christine Largeron, and Michel Beigbeder. 2019. Identifying sound descriptions in written documents. In 2019 13th International Conference on Research Challenges in Information Science (RCIS), pages 01–06. IEEE.

Claire Murphy. 2019. Olfactory and other sensory impairments in alzheimer disease. Nature Reviews Neurology, 15(1):11–24.

OpenAI. 2023. GPT-4 technical report. CoRR, arXiv.

Mylène Pardoen. 2019. Projet Bretez: une pincée de son dans l'Histoire. Digital Studies/Le champ numérique, 9(1):11.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Yael Ravin and Claudia Leacock. 2000. Polysemy: an overview. Polysemy: Theoretical and computational approaches, pages 1–29.

Benoît Sagot and Darja Fišer. 2012. Automatic extension of wolf. In GWC2012-6th International Global Wordnet Conference.

Flora Sakketou and Nicholas Ampazis. 2020. A constrained optimization algorithm for learning glove embeddings with semantic lexicons. Knowledge-Based Systems, 195:105628.

Richard Sorabji. 1971. Aristotle on demarcating the five senses. The Philosophical Review, 80(1):55–79.

Laura J Speed and Marc Brybaert. 2021. Dutch sensory modality norms. Behavior research methods, pages 1–13.

Serra Sinem Tekiroglu, Gözde Özbal, and Carlo Strapparava. 2014. Sensicon: An Automatically Constructed Sensorial Lexicon. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1511–1521, Doha, Qatar. Association for Computational Linguistics.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Lukas Vogel, Patrick Wonner, and Stefan M Huber. 2019. Chalcogen bonding: An overview. Angewandte Chemie International Edition, 58(7):1880–1891.

Cai Wingfield and Louise Connell. 2022. Sensorimotor distance: A grounded measure of semantic similarity for 800 million concept pairs. Behavior Research Methods.

Bodo Winter. 2019. Sensory linguistics: language, perception and metaphor. Converging Evidence in Language and Communication Research, 20.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. Explainability for large language models: A survey. arXiv preprint arXiv:2309.01029.

Yin Zhong, Mingyu Wan, Kathleen Ahrens, and Chu-Ren Huang. 2022. Sensorimotor norms for chinese nouns and their relationship with orthographic and semantic variables. Language, Cognition and Neuroscience, 37(8):1000–1022.

A Table of Notations

Table 3 sums up all notations used in the paper.

Notation	Definition
s	Sentence
$t(s)$	Tokenized sentence
$\mathbb{M}$	Ensemble of sensory functions, s.t. $\mathbb{M} = \{$OLF, GUS, AUD, VIS, HAP$\}$
m	Sensory function, s.t. $m \in \mathbb{M}$
D	Documents corpus
d	A document, s.t. $d \in D$
$D_{pos}(m)$	Subset of D containing all positive sentence examples w.r.t. the sensory function m
$D_{neg}(m)$	Subset of D containing all negative sentence examples w.r.t. the sensory function m
$C(s)$	Class label of sentence s (1 -positive- or 0 -negative-)
$\epsilon(s)$	Classification function for Step 1 of *SENSE-LM*)
w	Word, s.t. $w \in s$
$lemma(w)$	Lemma of word w
L_{SN}	Lancaster Sensorimotor Norms : Dictionary with words as keys and associated sensorimotor representations (11 dimensions) as values
w_{SN}	Sensory vector representation of the word w
s_{SN}	Sensory vector representation of the sentence s
$w_{SN}(j)$	jth dimension of w_{SN}
$w_{SN}(m)$	Dimension of w_{SN} associated to the sensory function m
s_B	BERT features extracted from the sentence s in Step 1
l	BERT's padding length
$F(w, m)$	Ground truth class label of word w w.r.t. the sensory function m, in Step 2
$\gamma(t(s), m)$	Classification function of *SENSE-LM*'s Step 2
$V(t(s), m)$	Vector output of $t(s)$ w.r.t the sensory function m in *SENSE-LM*'s Step 2
$\mathbb{L}_m$	Lexicon of terms related to the sensory function m
$P_{pos}(s, m)$	List of words predicted as positive in sentence s, w.r.t the sensory function m
w_{pos}	Word identified as positive, s.t. $w_{pos} \in P_{pos}(s, m)$
$P_{neg}(s, m)$	List of words predicted as negative in sentence s, w.r.t the sensory function m
$\mathbb{E}$	Set of semantic embeddings spaces
e	Semantic embedding space, s.t. $e \in \mathbb{E}$
T	Threshold value for semantic distances
U	Threshold value for sensorimotor dimension values
$\text{CosSim}_e(a, b)$	Cosine similarity between the representations of words a and b in the semantic space e
$Synsets(w)$	List of WordNet Synsets of word w

Table 3: Table of notations.

B Software and Hardware Setup

The experiments in this paper are executed using Python 3.10, PyTorch[7] version 1.13.1 and Keras for model architectures, NLTK (Bird et al., 2009) and SpaCy (Honnibal and Montani, 2017). Model pre-trained parameters are obtained from HuggingFace[8]. For the implementation of Step 2, we used and adapted an existing implementation[9]. The hardware environment in which experiments are conducted includes one NVIDIA RTX A5000 Mobile GPU (6144 CUDA Cores), one 11th Gen Intel® Core™ i9-11950H @ 2.60GHz × 16 CPU and 32 GB of RAM.

C Evaluation – Error Analysis

We provide the results of Step 2 for the Odeuropa dataset, grouped by Semantic Category in Table 4, for a more detailed reading of the actual performances of SENSE-LM. We note that SENSE-LM provides strong performances for the detection of Sensory Words, with a F1-Score over 90. It is expected as these words are most of the time explicit («odour, smell, perfume, etc...») and easy to identify as markers of odour, from the perspective of text features and sensorimotor features. However, SENSE-LM happens to struggle with Evoked Experiences; indeed, such expressions are few in number (only 5.8% of annotated terms) and do not always reflect explicitly the presence of an odour. It may be difficult to establish a semantic correlation with odours with too few examples.

Then, in Table 5, we provide the detailed results for the same scenario, grouped by Part-of-Speech:

Our model shows higher performances in particular for verbs and adjective. It is expected, as sensorimotor representations cover a wide spectrum of encountered words and verbs, providing strong assets on their relationship with an olfactory experience. It appears to show lower performances for Proper Nouns, that cannot be described from a sensorimotor point of view and may only be classified positively according to the text features. The model also struggles with numbers such as dates or counted entities; these are exception cases that are few in the dataset, which is a reasonable explanation on why we have difficulties to learn properly how to classify them.

[7] https://pytorch.org/

[8] https://huggingface.co/

[9] https://github.com/Jitendra-Dash/Extracting-Phrase-From-Sentence

D Evaluation of Computational Costs

In the following, we provide the costs of *SENSE-LM* compared to the baselines described in Section 4.3 and Section 4.4.

For each mechanism, we compare the number of model parameters, denoted **# Parameters**, the average duration of a single full model training over 5 trainings, denoted **Training Duration (s)**, and the average inference duration per data record, over all records of the test dataset, denoted **Inference Duration per record(s)**. The experiments are performed over the Odeuropa Benchmark dataset. The results for Step 1 are reported in Table 6, and the results for Step 2 in Table 7. The reported results correspond to experiments executed with the hardware setup described in Appendix B.

E Evaluation of Sensory Terms Extraction – Dataset Size Impact Analysis

In the following, we discuss the amount of labelled data required to benefit from the effective performances of *SENSE-LM* compared to baselines. In Figure 4, we plot the F1-Score of *SENSE-LM* and baselines according to the number of sentences labelled (i.e., sentences with an annotation of sensory terms), against the constant performances of Lexifield, that does not require preliminary data annotation. We plot the F1-Score on the Y axis, and the amount of labelled data on the X axis. For each point of the X axis, we incrementally augment the size of the dataset used to train the RoBERTa component in Step 2.1 of *SENSE-LM*. We observe that RoBERTa alone requires 80 labelled sentences to perform as good as Lexifield, while *SENSE-LM* is already better with only 10 sentences. However, we observe that it requires at least 300 labelled sentences to obtain a stable and optimal F1-Score. It is worth noting that our architecture remains better than RoBERTa in any case and acquires a stable behavior with fewer records, justifying the interest of Steps 2.2 and 2.3.

	# of groundtruth utterances	% of groundtruth utterances	Precision	Recall	F1-Score
Evoked Experience	196	5.8%	70.42 ± 2.41	52.03 ± 1.38	58.50 ±2.01
Quality	614	18.2%	75.41 ± 1.28	71.49 ± 2.01	73.40 ± 1.65
Sensory Source	1787	53.2%	71.11 ± 1.65	76.63 ± 1.87	73.66 ± 1.77
Sensory Word	764	22.7%	84.92 ±1.01	97.87 ±0.51	90.94 ± 0.72

Table 4: Evaluation of *SENSE-LM*'s sensory terms extraction step for Odeuropa, detailed by semantic category

	# of groundtruth utterances	% of groundtruth utterances	Precision	Recall	F1-Score
NOUN	1112	49.91 %	75.61 ± 1.22	74.62 ± 1.21	75.11 ± 1.22
ADJ	549	24.64 %	81.88 ± 1.56	73.82 ± 1.26	77.64 ± 1.41
VERB	261	11.71 %	83.33 ± 1.55	83.33 ± 1.71	83.33 ± 1.61
NUMBER	12	0.53%	65.23 ± 2.12	70.83 ± 2.41	67.91 ± 2.30
ADVERB	36	1.61%	80.55 ± 1.82	78.12 ± 1.18	79.31 ± 1.56
PROPER NOUN	258	11.57%	72.49 ± 1.34	74.22 ± 1.28	73.34 ± 1.31

Table 5: Evaluation of *SENSE-LM*'s sensory terms extraction step for Odeuropa, detailed by Part-of-Speech

	Odeuropa Benchmark Dataset			
Method	**# Parameters**	**Training Duration (s)**	**Inference Duration per record (µs)**	F1-Score
BERT	110M	336	239	90.80 ± 0.85
LR(s_{SN}))	22	2	11	76.97 ± 1.36
GPT-4	Over 100T	N/A	N/A	90.49 ± 1.61
SENSE-LM	110M	401	251	93.16 ± 0.76

Table 6: Evaluation of costs of *SENSE-LM*– Step 1 versus baselines.

	Odeuropa Benchmark Dataset			
Method	**# Parameters**	**Training Duration (s)**	**Inference Duration per record (ms)**	F1-Score
Lexifield ($\mathbb{L}_m$)	N/A	N/A	8	55.69 ± 1.25
GPT-4	Over 100T	N/A	N/A	60.62 ± 2.24
SENSE-LM (Step 2.1)	110M	377	18.12	72.52 ± 1.68
SENSE-LM (Step 2.1 ∪ Step 2.2)	110M	377	18.27	76.84 ± 1.74
SENSE-LM (Step 2.1 ∪ Step 2.3)	110M	377	19.67	74.99 ± 1.77
SENSE-LM (All steps)	110M	377	19.82	77.58 ± 1.65

Table 7: Evaluation of costs of *SENSE-LM*– Step 2 versus baselines.

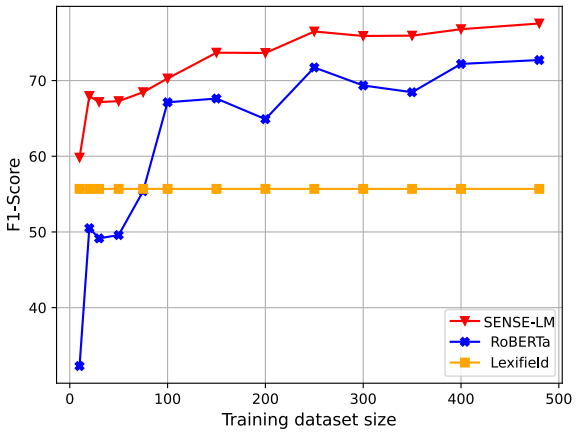

Figure 4: Training dataset size versus F1-Score trade-off for *SENSE-LM*'s Step 2, compared to baselines, for the Odeuropa dataset.

F GPT-4 Teaching protocols

We detail the protocols used to teach our different tasks to Chat GPT-4. We use the Chat GPT-4 web prompt[10]. We provide the detailed transcripts of the chat prompts corresponding to each task in our repository[11].

F.1 Auditory Dataset Generation

We provide the protocol used to generate the Auditory dataset that we described in 4.1. We askt GPT-4 generate 200 positive examples; i.e. auditory sentences, of length 10. Then, we generate 200 negative examples of length 10 as follows. We repeat the same protocol for 2 times 200 sentences *"between 25 and 35 words"*, and 2 times 100 sentences *"between 35 and 50 words"*, resulting in 1000 sentences. We check the consistence of the data manually; we corrected 11 misclassified sentences on 1000 generated examples. We did not notice any personal data, nor offensive content.

F.2 Binary Sentence Classification – GPT-4 Teaching Protocol

We provide the protocol used for teaching GPT-4 our binary classification task, as we consider it as a baseline with the objective of validating the relevance of our work, compared to the current capabilities of pre-trained models. We define the classification task as described in Section 4.3, we provide a set of examples corresponding to our training set to GPT-4, by providing both the sentences and their class (positive or negative), then we ask the model to classify the test set.

F.3 Sensory terms Extraction – GPT-4 Teaching Protocol

We use the same protocol described in Appendix F.2, applied to positive sentences only, by using the entire sentence as an input, and the set of words to be extracted as a target. We ask GPT-4 to predict the words to extract on the test set.

[10] https://chat.openai.com/

[11] https://github.com/cfboscher/sense-lm/tree/main/gpt4_prompts