# Are you not moved?
# Incorporating Sensorimotor Knowledge to Improve Metaphor Detection

**Ghadi Alnafesah**
University of Birmingham
Qassim University
gxa713@bham.ac.uk
gm.alnafesah@qu.edu.sa

**Phillip Smith**
University of Birmingham
p.smith.7@bham.ac.uk

**Mark Lee**
University of Birmingham
m.g.lee@bham.ac.uk

## Abstract

Metaphors use words from one domain of knowledge to describe another, which can make the meaning less clear and require human interpretation to understand. This makes it difficult for automated models to detect metaphorical usage. The objective of the experiments in the paper is to enhance the ability of deep learning models to detect metaphors automatically. This is achieved by using two elements of semantic richness, sensory experience, and body-object interaction, as the main lexical features, combined with the contextual information present in the metaphorical sentences. The tests were conducted using classification and sequence labeling models for metaphor detection on the three metaphorical corpora VUAMC, MOH-X, and TroFi. The sensory experience led to significant improvements in the classification and sequence labelling models across all datasets. The highest gains were seen on the VUAMC dataset: recall increased by 20.9%, F1 by 7.5% for the classification model, and Recall increased by 11.66% and F1 by 3.69% for the sequence labelling model. Body-object interaction also showed positive impact on the three datasets.

## 1 Introduction

Metaphors are an important and widespread form of language construction. A metaphorical sentence's meaning is not a direct, literal translation of its parts, but rather an overall collection of meanings in a specific context. For example, the phrase "*weigh my options*" refers to the situation in which the advantages and disadvantages of an option are examined for a decision. It is a CONSIDERATION, not a literal WEIGHING. This form of notation refers to the Conceptual Metaphor Theory (Lakoff and Johnson, 1980), where the source domain WEIGHING provides the words used to describe the target domain CONSIDERATION. Another example that discusses LOVE while addressing the concept of HEAT: "*I bumped into an old flame at the library*". Such examples demonstrate that understanding and interpreting metaphors are complex tasks for the field of NLP. Many fields, such as Information Extraction (Do Dinh et al., 2018; Le et al., 2020) and sentiment analysis (Rentoumi et al., 2012; Karanasou et al., 2015; Biddle et al., 2020) benefit from metaphor detection. Many experiments are being undertaken to improve the detection task using machine learning and deep learning models.

The term "sensorimotor knowledge" describes knowledge learned through the body's interactions with its surroundings. This knowledge could aid in comprehending metaphors, understanding how they are constructed, and consequently, enhance the performance of automated metaphor detection. By incorporating sensorimotor knowledge as a feature in neural network models, this improvement could become feasible. While some research has been conducted on sensory experience and conceptual norms for automated metaphor detection, as of the writing of this paper, no study on the impact of adding body-object interaction to neural network models as a feature for metaphor detection has been published.

The paper makes the following contributions:

1. The study aims to enhance the word/context representations provided by GloVe and ELMo vectors by incorporating scores from two datasets related to sensory experience and body-object interaction. These additional scores will serve as lexical features to improve the models' understanding of metaphors.

2. The study will conduct metaphor detection experiments using two different deep learning models. One model is designed for sentence-level metaphors and is based on the BiLSTM



*Proceedings of Recent Advances in Natural Language Processing*, pages 80–89
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_009

classification model proposed by (Gao et al., 2018). The other model is for word-level metaphors and relies on the sequence labeling model (RNN_HG) proposed by (Mao et al., 2019).

3. The performance of the two deep learning models will be evaluated on three corpora: VUAMC, TroFi, and MOH-X. These corpora likely contain diverse and varied examples of metaphors, which will provide insights into how well the models generalise across different datasets.

The paper is structured as follows: Section 2 provides a preview of the existing literature of related works. In Section 3, the theories forming the foundation of this study are introduced. Section 4 introduces the models and datasets used in the experiments, along with the steps to be followed in Section 5 for both models. Section 6 presents the analysis and decisions made during the experiments. Finally, Section 7 provides a conclusion, summarising the paper's findings.

## 2 Related Work

The concept of semantic richness comes from the theory of semantic representation, stating that information is stored and retrieved through an interconnected network of concepts. This network includes features and information contributing to the meaning of each concept (Pexman et al., 2007; Findlay and Carrol, 2018). Richer concepts have more semantic information, leading to faster activation, improved processing, and better decision-making in the brain (Kounios et al., 2009). Similar concepts may not evoke the same semantic information, showing varying levels of richness. Semantic richness is assessed based on two categories: elements related to the network's strength and elements linked to the perceptual aspect of the network (Findlay and Carrol, 2018). While numerous studies have examined strength-related elements in Natural Language Processing (NLP), like the number of features and neighborhood density (Pexman et al., 2002; Mason, 2004; Wilks et al., 2013; Goldberg, 2017), the elements associated with the perceptual part of the network have received less attention.

Based on shared information from the environment that senses sensory input (such as taste, sight, sound, etc.), language facilitates a common ground for communication. This idea holds true for both literal and figurative languages, as introduced by Tekiroğlu et al. (2015), who attempted to measure the impact of these sensorial elements on metaphor identification using a dependency-parsed corpus of adjective-noun (AN) pairs. Meanwhile, (Wan et al., 2020) tested the conceptual norms as a linguistic enhancement method for metaphor detection of VUAMC verbs. However, as of the date of this publication, the concept of body-object interaction has not been researched in association with automated metaphor detection. For the task of metaphor detection, in the hope of better automated detection, it is essential to understand this complex form of language, and these features could facilitate such understanding.

## 3 Theories

The mind is capable of forming mental images and evoking various sensations when reading or hearing certain words. This ability to trigger sensory and/or perceptual experiences in the mind is known as a sensory experience (Juhasz and Yap, 2013). For instance, when the word *incense* is encountered, the mind may generate a mental picture, and the word *fragrance* may evoke the actual smell associated with *incense*. Metaphors are a type of language that relies on describing a mental image to represent an abstract concept. They achieve this by using words from a concrete, sensed domain and applying them to another domain. As mentioned in the introduction, Lakoff and Johnson (1980) described the conceptual metaphor mapping where the concept of CONSIDERATION is depicted as a sensed WEIGHING experience. This theory is further developed in Lakoff et al. (1999), which suggests that bodily interactions with the environment are projected onto the new conceptual notions of these metaphors. This developed theory aims to explain how conceptual metaphors can be understood even when the direct experiential connection between the source and target domains is lacking, leading to some mappings being vague. For instance, the metaphor "*he is hungry for recognition*" can be understood by mapping DESIRE is HUNGER because "*food is desired*". This physical reaction of hunger is connected to the abstract idea of seeking recognition.

Based on the discussion above, metaphors create an image-scheme knowledge called the sensory interaction system, where abstract concepts

paint mental images of human bodily interactions. Words with a high score on the image scale could be used in metaphors where their score is noticeably higher than that of the context, further associated with some degree of human sensory input.

The body–object interaction rating reflects how easy it is to interact physically with the word's referent (Siakaluk et al., 2008). The high scale score indicates that the body's interaction with this concept is easier. For example, *key* was found to be more concrete and perceivable and linked to high-level sensory, haptic and visual experiences, this is in contrast to the word *mountains*, which scores low on the scale of body–object interaction with a lesser degree of the characteristics previously mentioned. One can see *mountains* but cannot interact with them with everyday human physical actions, while one can see, touch, and turn to unlock with *keys*. Within the cognitive sciences, researchers investigate the influence of body-object interaction measurement on various cognitive activities, including word recognition and information acquisition. The theory of the embodied view of cognition, as presented by (Siakaluk et al., 2008), posits that conceptual knowledge is grounded in perceptual interactions with the environment. This means that learning and understanding new concepts are built upon prior knowledge acquired through interactions with the surrounding environment. This observation could be applied to the idea of metaphors, as Lakoff et al. (1999)'s theory could be extended to the concept of body-object interaction. For instance, in the example '*this movie stinks*", it is clear that the statement expresses a negative remark about the movie, based on the known experience that '*stink is bad*". In other words, such metaphors can be easily comprehended because the tactile and visual experiences associated with them are akin to what the metaphor is referring to. Consequently, body-object interaction measures may aid in translating this knowledge into language that can be used to explain words and concepts. Particularly, abstract ideas could be better understood by employing this measure.

## 4 Metaphor Detection Experiments Setup

This section provides details about the experiments, which consist of two stages. The first stage is the preprocessing stage, where the SVM model trained on SEN and BOI lists assigns prediction scores to all tokens in the metaphorical corpora. In the second stage, the effect of these added predictions on metaphorical tokens is tested using the BiSTLM and RNN_HG models (Gao et al., 2018; Mao et al., 2019) for both sentence-level and word-level metaphor detection.

### 4.1 Metaphor Corpora and Other Datasets

Three metaphor corpora will be used; all will undergo the same steps from the preprocessing to the detection experiments. These datasets are the VUAMC, MOH-X and TroFi Gao et al. (2018) and Mao et al. (2019). The use of multiple datasets is essential to evaluate the performance of the models across various contexts and domains, ensuring that the models do not become overfitted to specific datasets and can generalise effectively to new data. Moreover, additional datasets related to sensory experience and body-object interaction will be used to train the SVM, which will be used to predict scores for all words in the mentioned metaphor corpora.

The VUAMC dataset (Tighe, 2010) is a manually annotated corpus containing metaphors from various registers. The MOH-X dataset (Mohammad et al., 2016) consists of simpler and shorter sentences compared to the other datasets, with each sentence having one labelled verb. Similarly, TroFi (Birke and Sarkar, 2006) shares similarities with MOH-X, having simpler sentences. The datasets utilised in the study contain lists of sentences, and the classification datasets (VUAMV, MOH-X, and TroFi) are labeled as 0 for literal and 1 for metaphor, based on the presence of a metaphorical verb. On the other hand, the VUAMC sequence dataset labels each word in the sentence for metaphoricity. The sequence model utilises the MOH-X and TroFi datasets, using 1 for the target verb if it is a metaphor, and 0 for all other words in the sentence.

Juhasz and Yap (2013) published a 5,000-word English word list rated for their sensory experience. The words were rated on a scale of 7, where low numbers indicated a low image/sensory impact. For example, the word *intent* was assigned a sensory experience rating of 2.40, while *balloon* received a rating of 5.45, indicating a richer image/sensory impact. The scoring scale was later reorganized as integers, resulting in rating results ranging from 1 to 6. In their study, Pexman et al. (2019) compiled a word list containing over 9,000 English words rated for their *ease of body interaction* on a scale of 1 to 7. A very low score signifies that it is challenging

| Dataset | Total | Meta. | Lit. |
|---|---|---|---|
| VUAMC CL | 10,489 | 2,837 | 7,652 |
| VUAMV SQ | 17,932 | 4,717 | 16,064 |
| MOH-X | 214 | 192 | 195 |
| TroFi | 3737 | 50 | 50 |

Table 1: The breakdown of the metaphorical datasets, number of sentence, tokens, metaphor and literal.

| Dataset | Total | SEN | BOI |
|---|---|---|---|
| SEN | 5856 | _ | _ |
| BOI | 9349 | _ | _ |
| VUAMC Cl. | 17017 | 3059 | 3572 |
| VUAMC Seq. | 16979 | 3156 | 3701 |
| MOH-X | 1677 | 694 | 811 |
| TroFi | 13738 | 2771 | 3216 |

Table 2: The breakdown of total tokens for each dataset and the number that the sensory experience and body–object interaction list covers.

for the body to interact with those words physically. For instance, the word *ceiling* has a low score of 2.5 because it is not easy to perform physical actions like jump and touch with the *ceiling*. In contrast, the word *chair* received a higher rating on the scale, 6.88, indicating that it is easy to interact physically with it; one can underlinetouch, move and sit on a *chair* with ease. The scale was later rearranged as integers, ranging from 1 to 6.

Table 1 provides an overall statistics about the metaphorical datasets. The VUAMC CL utilised in the classification experiments contains a total of 10,489 sentences, out of which 2,837 are metaphorical, and 7,652 are literal. In the sequence experiments, the VUAMC SQ comprises 15,820 sentences with 17,932 tokens. Among these, 4,717 tokens are metaphors, and 16,064 tokens are literal. The MOH-X dataset includes 647 sentences with 214 unique target verb tokens, 192 appear in metaphorical sentences, and 195 in literal sentences. Lastly, the TroFi dataset consists of 3,737 sentences with 50 unique verb tokens. Each of these verb tokens is found in both metaphorical and literal sentences. In addition, Table 2 displays the token count for each dataset and indicates how many tokens are covered by the sensory and body-object lists. These lists will be utilised to train the SVM, enabling the assignment of predicted sensory and body-object scores to each token in the metaphorical dataset. The original ratings and the predicted ratings will be evaluated separately during the metaphor detection process.

### 4.2 Embeddings

In the preprocessing step, the SVM utilises BERT pre-trained as the vector representation for the words in the datasets. GloVe and ELMo are used only in the metaphor detection experiments as the word/context representations for VUAMC, MOH-X and TroFi.

GloVe is a 300-dimensional word embedding for word meaning derived from statistical techniques used to calculate word–context co-occurrence in a large corpus (Pennington et al., 2014). Embeddings from Language Models (ELMo) (Peters et al., 2018) are deep 1,024-dimensional contextualised embeddings that represent each word's whole sentence input, where the context of a word's surroundings is considered, resulting in a dynamic representation that can change based on the sentence in which it appears. Along with GloVe, they make a 1,324-dimensional vector that represents each word in the sentence for the three metaphor datasets.

The Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) are formed by deep bidirectional representations that capture the contextual information from both the right and left sides of unlabelled text. Each word in the sensory and body-object dataset word list is assigned a BERT representation serving as SVM inputs during training. Similarly, every word in the VUAMC, MOH-X, and TroFi acquires a BERT representation and receives predicted scores from the SVMs. These predicted scores will be tested in the subsequent metaphor detection stage.

### 4.3 Models

The baseline[1] used to evaluate the sentence-level metaphor detection was built on the BiLSTM proposed by Gao et al. (2018). The word-level metaphors were tested using the RNN_HG sequence labelling model introduced by Mao et al. (2019). Both models rely on GloVe and ELMo embeddings to capture contextual information. The experiments on VUAMC use three splits for training, validation, and testing, whereas the tenfold

[1]The authors wrote in the README file that running the provided script is expected to result in some numbers that are **lower** than the reported numbers because the reported numbers in the paper were achieved with early stopping and additional training with smaller learning rates. These details were not included in the available scripts and were not provided in the paper.

cross-validation technique is applied for MOH-X and TroFi. Table 3 presents the hyperparameters used for these baselines.

### 4.4 Preprocessing Stage

The main objective of this paper is to assess the impact of enhanced contextual information on the metaphor detection task by incorporating sensory and body-object scores. Additionally, the study aims to gain deeper insights into the relationship between these elements and metaphors. The testing of the effect of sensory and body-object scores on the metaphor detection task involves three steps. Section 5, describes these three steps in detail for VUAMC, MOH-X, and TroFi.

1. Test the metaphor detection models when VUAMC, MOH-X, and TroFi words are present in the sensory and body-object datasets. Out-of-list words are assigned a score of 0. The main aim is to assess the impact of incorporating the current ratings of the sensory and body-object datasets on metaphor detection, without relying on pre-trained tools for prediction.

2. Each word in the sensory and body-object datasets receives a pre-trained BERT embedding, which is then used as input to train two SVMs: one for sensory and another for body-object. Next, each word in VUAMC, MOH-X, and TroFi datasets is assigned a pre-trained BERT embedding, allowing the SVMs to provide single-digit sensory and body-object scores. These obtained scores are then combined with GloVe and ELMo embeddings, creating input for the classification and sequence labelling models used in metaphor detection.

3. Similar to step 2 for training the SVM, however, the SVM assigns the predicted scores as probability distributions with six digits. Each digit represents the probability that the word falls under a specific score. These digits are then concatenated with GloVe-ELMo embeddings to create input for the metaphor detection models. The objective of this step is to evaluate the value of using higher-detail scores as probabilities, in contrast to single-digit scores.

## 5 Implementation and Results

### 5.1 SVMs

In steps 2 and 3 of the preprocessing, the SVM is employed to assign sensory and body-object predicted scores to all words in the VUAMC, MOH-X, and TroFi datasets. Initially, the SVM is trained on the sensory and body-object datasets using BERT pre-trained vectors as input to represent each word in these lists. Next, BERT pre-trained vectors are extracted for each word in the metaphorical corpora, and these vectors are then utilized in the SVM to assign a predicted score for each word.

Subsequently, the predicted scores are concatenated with the GloVe-ELMo embeddings to serve as the input for the metaphor detection models. As reported by Alnafesah et al. (2020), integrating a probability distribution for concreteness rating into both the classification and sequence labelling models yielded significant improvements in performance, with the F1 scores increasing by 10.23% and 6.81%, respectively. The probability distribution provided valuable information regarding the scoring of specific words, leading to improved performance in metaphor detection for the models. Table 4 displays the F1 scores obtained from the tenfold cross-validation during the training of the SVMs on the sensory and body-object datasets.

### 5.2 Classification and Sequence Labelling for Metaphor Detection

The sentence-level metaphor detection baseline is established using the BiSTLM model introduced by Gao et al. (2018). This model classifies sentences as either *metaphor* (assigning 1) or *literal* (assigning 0) based on the target verb and its surrounding context. For word-level metaphor detection, the baseline is built on the RNN_HG model proposed by Mao et al. (2019). This model assigns a label of 1 for *metaphor* or a label of 0 for *literal* to each word in the sentence, based on the word's surrounding context. This section presents the results for each step of testing the sensory experience and body-object interaction using these models for metaphor detection task.

In step 1, only words in VUAMC, MOH-X, and TroFi that are present in the sensory and body-object datasets are given scores. This step aims to evaluate the existing ratings without the intervention of the SVM. Table 5 displays the results of this step, denoted as *SEN_ST1* and *BOI_ST1*, for precision, recall, and F1 in both detection models.

| Exp. | P. | R. | F1 | H_size | Drop1 | Drop3 | B_size | Layer | Epch |
|---|---|---|---|---|---|---|---|---|---|
| **VUAMC Cl.** | 56.28% | 51.107% | 53.57% | 128 | 0.3 | 0.2 | 64 | 1 | 20 |
| **VUAMC Seq.** | 76.23% | 64.45% | 71.22% | 256 | 0.5 | 0.2 | 2 | 1 | 29 |
| **MOH-X Cl.** | 75.44% | 77.393% | 76% | 300 | 0.2 | 0.2 | 10 | 1 | 30 |
| **MOH-X Seq.** | 76.408% | 81.63% | 78.46% | 256 | 0.5 | 0.1 | 2 | 1 | 20 |
| **TroFi Cl.** | 69.661% | 73.04% | 71.088% | 300 | 0.2 | 0 | 10 | 1 | 15 |
| **TroFi Seq.** | 68.98% | 74.489% | 71.575% | 256 | 0.5 | 0.1 | 2 | 1 | 20 |

Table 3: The hyperparameters used to acquire the baselines used for these experiments. The classification experiments are built on Gao et al. (2018), and the sequence labelling experiments are built on Mao et al. (2019). Precision, Recall and F1 for MOH-X and TroFi are the best of the tenfold cross-validation.

| SVM | F1 Mean | F1 MAX | F1 MIN |
|---|---|---|---|
| SEN-single | 38.678% | 43.247% | 35.213% |
| SEN-prob. | 40.026% | 44.273% | 37.606% |
| BOI-single | 38.881% | 42.887% | 36.791% |
| BOI-prob | 39.03% | 41.711% | 35.508% |

Table 4: The mean, max. and min. F1 scores for each SVM trained on the sensory experience and body–object interaction datasets.

The F1 scores for sensory with the classification model showed an increase in all three datasets, with VUAMC experiencing the highest increase, reaching 57.603% from 53.57%. Similarly, recall for VUAMC increased, reaching 61.839%. However, precision decreased in VUAMC and TroFi, while there was a small increase of 1.67% in MOH-X. As for body-object, precision increased for VUAMC and MOH-X, while recall and F1 for MOH-X and TroFi showed the opposite trend. TroFi's recall showed a greater increase, reaching 76.613%, while MOH-X's F1 showed a better increase, reaching 77.048%. The sequence model with sensory showed improvement in F1 for all three datasets. Recall increased in VUAMC (71.199%) and TroFi (76.471%), but slightly decreased in MOH-X. Similarly, for body-object, F1 increased in all datasets. Recall in VUAMC reached 73.298% and 75.638% for TroFi, while it decreased slightly in MOH-X.

In step 2, the SVMs' single-score assigned predictions are tested. These models assign sensory and body-object scores as a single digit to all words in the three metaphor datasets. The results are shown under *SEN_ST2* and *BOI_ST2* in Table 5. The classification with sensory experiments in VUAMC, there were minimal changes in all three metrics. Recall and F1 of the MOH-X and TroFi datasets increased slightly, while precision decreased slightly. For body-object in the MOH-X dataset, there were increases in all three metrics. On the other hand, in VUAMC's results, precision increased to 58.04%, while recall and F1 decreased. In contrast, recall and F1 increased for TroFi dataset, while precision decreased. The TroFi dataset experiments with sensory showed improvement in all three metrics. F1 for VUAMC and MOH-X increased to 73.85% and 79.033%, up from 71.22% and 78.46%, respectively. Precision decreased in VUAMC and increased in MOH-X, while the opposite was true for recall in both datasets. For the body-object in the sequence experiments, there was an overall increase in almost all metrics for all datasets. However, precision for VUAMC and TroFi showed decreases in the results.

In step 3, the sensory and body-object predictions as a probability distribution are tested for all three datasets. The predictions, in the form of a six-digit score, are added to the vectors for all words. The results are shown in Table 5 under *SEN_ST3* and *BOI_ST3*. The classification model for MOH-X with the sensory experiments showed an increase in results for all three metrics. VUAMC's and TroFi's recall and F1 increased, while precision slightly decreased. VUAMC's F1 reached 56.1%, and recall reached 56.7%. However, for body-object, the model's performance with VUAMC showed a decrease in all metrics. On the other hand, precision and F1 increased in MOH-X, while recall decreased very slightly from 77.393% to 77.358%. As for TroFi, F1 and recall increased to 75.418% and 71.664%, respectively, but Precision decreased to 68.416%. For MOH-X and TroFi, all three metrics increased slightly in the sensory experiments with the sequence model. In contrast, precision decreased in VUAMC, while recall and F1 increased to 71.968% and 73.27%, respectively. Similarly, in body-object, VUAMC's precision decreased to 74.96%, while recall increased to 70.799%, and F1 to 72.82%. However, recall decreased in MOH-X and TroFi, while F1 increased to 79.275% and 71.925%, respectively. Precision also increased in

| Classification | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Exp.** | **Metrics** | **Baseline** | **SEN_ST1** | **SEN_ST2** | **SEN_ST3** | **BOI_ST1** | **BOI_ST2** | **BOI_ST3** |
| | **P.** | 56.28% | 53.91% | 56.13% | 55.4% | 63.06% | 58.04% | 56.85% |
| **VUAMC** | **R.** | 51.107% | **61.839%** | 50.65% | 56.7% | 33.84% | 49.17% | 48.2% |
| | **F1** | 53.57% | **57.603%** | 53.25% | 56.1% | 44.05% | 53.24% | 52.21% |
| | **P.** | 75.44% | 76.7% | 74.115% | 76.622% | 77.566% | 76.43% | 77.272% |
| **MOH-X** | **R.** | 77.393% | **80.209%** | 78.943% | 77.328% | 77.655% | **78.39%** | 77.358% |
| | **F1** | 76% | **77.807**% | 76.216% | 76.468% | 77.048% | **77.097**% | 76.88% |
| | **P.** | 69.661% | 68.937% | 69.068% | 68.852% | 68.18% | 68.439% | 68.416% |
| **TroFi** | **R.** | 73.04% | 73.952% | **75.961%** | 74.113% | **76.613%** | 75.216% | 75.418% |
| | **F1** | 71.088% | 71.312% | **72.159**% | 71.227% | **71.98%** | 71.643% | 71.664% |
| **Sequence Labelling** | | | | | | | | |
| **Exp.** | **Metrics** | **Baseline** | **SEN_ST1** | **SEN_ST2** | **SEN_ST3** | **BOI_ST1** | **BOI_ST2** | **BOI_ST3** |
| | **P.** | 79.58% | 76.23% | 78.1% | 74.6% | 75.188% | 76.84% | 74.96% |
| **VUAMC** | **R.** | 64.45% | 71.199% | 70.046% | **71.968%** | **73.298%** | 69.46% | 70.799% |
| | **F1** | 71.22% | 73.629% | **73.85%** | 73.27% | **74.23%** | 72.97% | 72.82% |
| | **P.** | 76.408% | 78.82% | 78.795% | 77.562% | 78.403% | 77.396% | 79.439% |
| **MOH-X** | **R.** | 81.63% | 80.053% | 79.809% | **81.804%** | 80.292% | **83.192%** | 79.841% |
| | **F1** | 78.46% | **79.257%** | 79.033% | 79.204% | 78.967% | **79.767%** | 79.275% |
| | **P.** | 68.98% | 68.598% | 69.358% | 69.963% | 68.984% | 67.579% | 70.03% |
| **TroFi** | **R.** | 74.489% | **76.471%** | 75.591% | 74.439% | 75.638% | **77.311%** | 74.1255% |
| | **F1** | 71.575% | 72.172% | **72.212%** | 72.001% | **72.066%** | 71.99% | 71.925% |

Table 5: The results of the classification and sequence labelling experiments for both sensory and body-object are presented. The results for MOH-X and TroFi represent the best performance from the tenfold cross-validation. The highest values of recall and F1 for each dataset under each feature are highlighted in bold.

both MOH-X and TroFi to 79.439% and 70.03%, respectively.

## 6 Analysis and Discussion

When analysing the incorrectly predicted files for the classification model of sensory experience, prepositions frequently appear, and the word *get* is prominent on the list. For instance, in the sentence "*probably need to get Ken's permission!*", all words received a sensory predicted score of 1, except for the word *Ken*, which received a score of 2. Despite the slight shift in ratings, especially for the word *Ken* following the target verb *get*, the model failed to make the correct prediction. This could be attributed to the very low sensory experience ratings for all words in the context, along with the nature of the word *get* as a metaphor. Words like *get* and others in similar situations are frequently used words that have lost their metaphorical meaning and have become literal. Additionally, the misprediction could be due to the lack of a noticeable rating shift, causing the model to overlook the metaphoricity indicators.

In another example, the case of the phrasal verb appears in "*I'll get some tables up with erm*" where the SVM's predicted sensory ratings for "*get some tables up*" were 1, 1, 3, and 2. The words *get* (1) and *up* (2) had slightly shifted ratings. However, it is possible that the model did not detect the phrasal verb due to the distance between its parts. Additionally, the model's decision could be related to the actual meaning of the sentence. In this context, *get* is used as the literal verb *acquire*, and the word *table* represents an *object that can be acquired* in a literal sense. Because there was no significant shift in the sensory ratings with the rating distance, the model classified the sentence literally based on these factors.

The word *produce* in the sensory rating was also misclassified in the sentence "*He chuckled, produced two cardboard cups, and poured me a generous slug of the whiskey.*" A similar situation was observed where the meaning of the target word matched the context, and there were no noticeable rating shifts. As a result, the model incorrectly classified the sentence as literal. In this case, the phrase "*produced two cardboard cups*" could be interpreted as literal since it refers to the actual action of creating the object *cup*. However, the intended meaning of the sentence was likely *bring out two cups* rather than *make two cups*. The phrase *to produce* can also mean *to bring out* or *to make apparent or present to the public*. This alternative meaning is what the sentence is likely trying to convey. The lack of noticeable sensory experience rating shifts (with sensory ratings of 2, 1, 3,

and 3) might not have provided enough helpful information for the classification model to correctly understand the intended meaning of the sentence. As a result, it classified the sentence as literal based on the available information.

The analysis of the incorrectly predicted files for body-object with classification reveals that the words *got* (body-object rating of 1), *go* (body-object rating of 2), and *back* (body-object rating of 5) appear at the top of the list. For instance, in the example "*having got some of the plumbing details wrong*" (*having* 2, *got* 1, *some* 1, *of* 1, *the* 1, *plumbing* 3, and *details* 2), it is evident that the rating shift between the words is slight, from 1 to 2 to 3. Furthermore, the words *plumbing* (rating of 3) and *details* (rating of 2) do not indicate strong physical manipulation. These factors combined could explain why the model's performance did not exhibit significant improvements, as observed in the sensory experiments. The same reasoning applies to the sentence "*Well, hang on a minute,*" where the ratings are 1, 1, 2, 1, and 1, with the verb *hang* as the target. The notable shifts in ratings could not provide any additional information beyond what was already known from the GloVe and ELMo embeddings.

The sensory ratings with sequence experiments misclassified the word *plant* in the sentence "*pull all nuclear plant out of the impending sale.*" as a non-metaphor. When examining the sensory ratings (*pull* 3, *all* 1, *nuclear* 3, *plant* 3, *out* 2, *of* 1, *the* 1, *impending* 2, and *sale* 2), the ratings were low, combined with the lack of an apparent shift, which may have led to missing the metaphoricity hints in using the word plant with *nuclear*. The word *down* in "*two dressing rooms and toilets down there.*" could be explained in the same way. The lack of noticeable rating shifts (*two* 1, *dressing* 2, *rooms* 3, *and* 1, *toilets* 3, *down* 2, and *there* 2) and the matching meaning of the word *down* as the direction, with the context being the *position of rooms*, could have pushed the model to misclassify the word *down* as literal.

The words *got* and *go* are also among the incorrectly predicted words for the sequence labelling experiments. The body–object interaction rating for *got* is 1 and for *go* is 2, both of which have low body–object interaction ratings. In the example "*I 've only got until tomorrow.*" the model misclassified the word *got* as literal. The body–object interaction ratings (*I* 2, *'ve* 1, *only* 1, *got* 1, *until* 1, and *tomorrow* 2) show no noticeable shifts between the words. Although the word tomorrow indicates time, and *got* indicates a somewhat physical action, the model should not misclassify the word *got*, because handling time physically is impossible. However, the body–object interaction rating did not reflect that when it gave the word *got* a low body–object interaction rating.

According to Pexman et al. (2019), the ratings reflect the ease of physical interaction with these words. Some of the words, in their sense, are similar; however, their body-object interaction ratings are different. For instance, *he* has a body-object interaction rating of 2.96, *she* has 3.30, *boy* has 4.9, and *girl* has 5.52. These variations in ratings for words that are supposed to be close in meaning space could have affected the metaphor learning with the body-object interaction ratings. Furthermore, Pexman et al. (2019) stated that the ratings were derived from concreteness and imageability ratings, along with other variables, but these specific variables were not specified.

## 7 Conclusion

This paper examined the impact of adding sensorimotor knowledge (sensory experience and body-object interaction) as external lexical resources to neural network models for the automatic detection of metaphors in text. Sensory relies on an image scheme, where a mental image is evoked along with other sensory activations to convey the intended meaning. On the other hand, body-object is derived from concreteness and imageability, which describe how easy it is to interact with a particular entity. Both concepts have been extensively studied in fields outside of NLP. The ratings from the lists, as well as the ratings obtained from trained SVMs, were tested on three metaphorical datasets using two types of deep learning models: one classifies sentences, and the other classifies words as literals or metaphors based on context. The models' performances demonstrated promising results, showing improvements in recall and F1 for metaphor detection across the three datasets. For future work, a more comprehensive breakdown of the variables could be used to acquire the ratings for sensory and body-object lists. This could include making the type of activated sensory more apparent. Additionally, considering imageability and concreteness in these tests might help bridge the gap in some of the variations in ratings observed, as mentioned

previously (he and boy, girl and she).

## References


Ghadi Alnafesah, Harish Tayyar Madabushi, and Mark Lee. 2020. Augmenting neural metaphor detection with concreteness. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 204–210.

Rhys Biddle, Aditya Joshi, Shaowu Liu, Cecile Paris, and Guandong Xu. 2020. Leveraging sentiment distributions to distinguish figurative from literal health reports on twitter. In *Proceedings of The Web Conference 2020*, pages 1217–1227.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 329–336.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Erik-Lân Do Dinh, Steffen Eger, and Iryna Gurevych. 2018. Killing four birds with two stones: Multi-task learning for non-literal language detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1558–1569, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Holly Findlay and Gareth Carrol. 2018. Contributions of semantic richness to the processing of idioms. *The Mental Lexicon*, 13(3):311–332.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. *arXiv preprint arXiv:1808.09653*.

Yoav Goldberg. 2017. Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, 10(1):1–309.

Barbara J Juhasz and Melvin J Yap. 2013. Sensory experience ratings for over 5,000 mono-and disyllabic words. *Behavior Research Methods*, 45(1):160–168.

Maria Karanasou, Christos Doulkeridis, and Maria Halkidi. 2015. Dsunipi: An svm-based approach for sentiment analysis of figurative language on twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 709–713.

John Kounios, Deborah L Green, Lisa Payne, Jessica I Fleck, Ray Grondin, and Ken McRae. 2009. Semantic richness and the activation of concepts in semantic memory: Evidence from event-related potentials. *Brain research*, 1282:95–102.

George Lakoff and Mark Johnson. 1980. Conceptual metaphor in everyday language. *The Journal of Philosophy*, 77(8):453–486.

George Lakoff, Mark Johnson, and John F Sowa. 1999. Review of philosophy in the flesh: The embodied mind and its challenge to western thought. *Computational Linguistics*, 25(4):631–634.

Duong Le, My Thai, and Thien Nguyen. 2020. Multi-task learning for metaphor detection with graph convolutional neural networks and word sense disambiguation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8139–8146.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3898.

Zachary J Mason. 2004. Cormet: a computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, 30(1):23–44.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33.

Jeffrey Pennington, Richard Socher, and Christopher D 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.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Penny M Pexman, Ian S Hargreaves, Jodi D. Edwards, Luke C. Henry, and Bradley G. Goodyear. 2007. The neural consequences of semantic richness. *Psychological Science*, 18(5):401–406. PMID: 17576279.

Penny M Pexman, Stephen J Lupker, and Yasushi Hino. 2002. The impact of feedback semantics in visual word recognition: Number-of-features effects in lexical decision and naming tasks. *Psychonomic bulletin & review*, 9(3):542–549.

Penny M Pexman, Emiko J Muraki, David M Sidhu, Paul D Siakaluk, and Melvin J Yap. 2019. Quantifying sensorimotor experience: Body–object interaction ratings for more than 9,000 english words. *Behavior research methods*, 51(2):453–466.

Vassiliki Rentoumi, George A Vouros, Vangelis Karkaletsis, and Amalia Moser. 2012. Investigating metaphorical language in sentiment analysis: A sense-to-sentiment perspective. *ACM Transactions on Speech and Language Processing (TSLP)*, 9(3):1–31.

Paul D Siakaluk, Penny M Pexman, Laura Aguilera, William J Owen, and Christopher R Sears. 2008. Evidence for the activation of sensorimotor information during visual word recognition: The body–object interaction effect. *Cognition*, 106(1):433–443.

Serra Sinem Tekiroğlu, Gözde Özbal, and Carlo Strapparava. 2015. Exploring sensorial features for metaphor identification.

Brian Tighe. 2010. Vu amsterdam metaphor corpus online.

Mingyu Wan, Baixi Xing, Qi Su, Pengyuan Liu, and Chu-Ren Huang. 2020. Sensorimotor enhanced neural network for metaphor detection. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 312–317.

Yorick Wilks, Adam Dalton, James Allen, and Lucian Galescu. 2013. Automatic metaphor detection using large-scale lexical resources and conventional metaphor extraction. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 36–44.