# Decoding Neural Representations of Sentences in Individuals with Autism

**Elaheh Toulabinejad**
Department of Computing Science
University of Alberta
Edmonton, Alberta
toulabin@ualberta.ca

## Abstract

Autism is a neurodevelopmental disorder characterized by challenges in social communication. Some of these challenges may arise from specific language profiles of autistic individuals. Research has shown that autistic individuals often struggle with interpreting figurative language. Language comprehension in the brain involves complex processes. One effective way to gain insights into these processes is by using machine learning techniques to decode neural activity. Natural language processing methods have shown promise in this area. In this study, we want to understand if the representations in the brain align with the semantic representations in a language model for various types of sentences among autistic individuals. Second, we seek to compare the representations of metaphors between autistic and non-autistic participants to determine if they are similar or different. To achieve this goal, we used classification and decoding of neural activity while participants processed metaphors. By addressing these questions, we hope to gain insights into how autistic individuals process language differently compared to non-autistic individuals, particularly in relation to figurative language understanding.

## 1 Introduction and Related Works

This paper explores how individuals with Autism Spectrum Disorder (ASD) process metaphors by examining their linguistic characteristics using functional Magnetic Resonance Imaging (fMRI) insights. By integrating Machine Learning and Natural Language Processing into neuroimaging data analysis, we aim to understand how these individuals interpret language, particularly figurative language. This enhances our understanding of how language processing works in the human brain.

### 1.1 Autism Spectrum Disorder (ASD)

Autism is a neurodevelopmental disorder influenced by genetic and environmental factors that affect the development of the brain. Autism can manifest in a range of severities, from severe forms to milder forms. Individuals with autism commonly demonstrate restricted and repetitive patterns of behavior, characterized by difficulties in communicating and interacting socially with others [1, 2].

### 1.2 Language and Speech Characteristics in Autism

Social skills and language development are closely connected. A common challenge for individuals with autism is difficulty with social communication. Developing language skills is crucial for improving social abilities, and improving social skills can also support language development. Therefore, there has been extensive study into the language profiles of individuals with autism [3].

Individuals with autism often demonstrate difficulties in semantics, syntax, and pragmatics aspects of language [3, 4]. They struggle with producing complex grammatical structures and understanding complicated syntax [5, 6]. In terms of semantic difficulties, they may struggle with semantic integration, using appropriate vocabulary, and understanding figurative language as they often interpret sentences very literally [3, 7].

To understand figurative language, it should be processed in three steps: Access, Integration, and Selection [8, 9]. The first step in understanding metaphors involves gathering all the necessary information from each word in the sentence. This information is then combined to create both the literal and nonliteral interpretations of the sentence. Next, the intended meaning must be chosen, which involves suppressing or ignoring irrelevant information or meanings that are not intended [10].

### 1.3 Functional Magnetic Resonance Imaging (fMRI)

Functional Magnetic Resonance Imaging (fMRI) has become a vital tool in neuroimaging. This non-invasive method tracks brain activity by detecting changes in blood oxygen levels. fMRI signals indirectly show neuronal activity, offering insights into how brain regions are active during specific cognitive tasks.

### 1.4 Machine Learning in Neuroimaging Data analysis

Multivariate Pattern Analysis (MVPA) is a collection of methods used to study neural responses by looking at patterns of activity in the brain. This allows researchers to explore the different brain states that a particular brain area or system can generate, which in turn increases our ability to decode information from brain activity. MVPA helps to extract detailed information from brain signals, enhancing our understanding of how the brain functions under different conditions or during specific tasks [11]. Using machine learning methods to analyze brain imaging data is a common practice in this field. Tasks such as decoding and encoding are well-known in this context [12].

Decoding involves predicting the stimulus (or another variable) based on neural activity. On the other hand, encoding involves predicting neural activity from stimuli. These concepts are fundamental in analyzing brain activity and understanding how stimuli are represented in neural patterns [13, 14]. Before, researchers have tried to understand how people with autism and those without autism understand metaphors [9, 8, 10].

## 2 Data

The data used in this study was collected by Chouinard and Cummine in 2016 [9]. The purpose of their study was to evaluate whether the processing of metaphorical and non-metaphorical sentences in individuals with ASD is the same as or different from individuals without ASD. In this section, we provide details about the dataset.

### 2.1 Stimuli

The set of sentence stimuli in the dataset contains 160 sentences. All the sentences follow the format *"Some x are y,"* where $x$ represents a category and $y$ is a known example of that category. Based on the words used instead of $x$ and $y$, these sentences can be grouped into four categories: 80 Literally True (LT) sentences, 40 Literally False (LF) sentences, 20 Metaphors (M), and 20 Scrambled Metaphors (SM).

In Literally True sentences, $y$ actually belongs to the category $x$. These sentences only have a literal meaning. Metaphors have both a false literal and a true nonliteral meaning. Literally False and Scrambled Metaphor sentences do not have a literal or nonliteral meaning, but there is a key difference between them. Literally False sentences were constructed by scrambling the Literally True sentences. Scrambled Metaphor sentences were made up of the same words as the Metaphor sentences. Table 1 contains examples from each category.

Table 1: Example stimuli for each category.

| Category | Sentence |
|---|---|
| LT | Some birds are robins |
| | Some clothes are pants |
| | Some fabrics are silk |
| LF | Some dances are lawyers |
| | Some fabrics are verbs |
| | Some fish are sugary |
| M | Some hands are magic |
| | Some ideas are diamonds |
| | Some jobs are jails |
| SM | Some desks are diamonds |
| | Some hands are jungles |
| | Some ideas are snakes |

### 2.2 Participants

Twelve participants with autism and twelve participants without autism took part in the data acquisition experiment. English was their first language, and all of them passed a hearing test. Both groups were similar in several important ways: they had comparable chronological age, non-verbal IQ, language abilities, and handedness. This ensured a fair comparison between the groups in the study.

Participants were excluded from the study if they had previous neurological injury, a history of hearing loss, contraindications for MRI scanning (such as claustrophobia or having metal in the body), or if they disclosed a psychiatric disorder.

### 2.3 Procedure

The participants were tested individually while inside a 1.5 T MRI scanner. They were informed that they would hear a sentence and needed to judge whether the sentence was literally true or false. There was no mention of the possibility of encountering metaphorical stimuli. All participants used their right hand to indicate "true" or "false" using the MRI response keys and received explicit instructions to respond as quickly and accurately as possible. Responses and response times were recorded.

### 2.4 Reprocessing

Preprocessing was conducted using SPM 8 and involved several steps. First, they realigned images from both runs to align brain volumes and correct for motion artifacts in the scans. Additionally, fMRI scans were co-registered to the participant's anatomical brain scan. Then, the brain was segmented into gray matter, white matter, and cerebrospinal fluid maps, and forward and inverse deformation fields were generated. Using the results of the segmentation step, the participant's anatomical data was transformed into the Montreal Neurological Institute (MNI) space, and regions of interest (ROI) masks were applied to extract voxels from specific regions. Finally, the inverse deformation fields were used to revert the participant's ROI scans back to their original anatomical space.

To improve the signal-to-noise ratio (SNR), the functional fMRI data were pre-processed to calculate a beta weight for each stimulus sentence. Beta weights represent estimates of the best fit of a canonical BOLD response to the neural signal. They ran a General Linear Model (GLM) on the pre-processed fMRI data for each participant to obtain their beta weights.

We selected data from specific ROIs known to be important for language processing and decision-making. These regions include the left angular gyrus (AG) for semantic processing, left inferior frontal gyrus (IFG) for phonological and semantic processing, and left dorsal anterior cingulate cortex (dACC) for decision-making. We combined the voxel data from these regions to create a larger region called "Language-ROI" [10].

## 3 Methods

Our main goal is to identify patterns and connections related to metaphor processing in brain signals. To achieve this, we employ machine learning methods to extract insights from the dataset mentioned in Section 2. In this section, we explain our experiments. [1]

### 3.1 Beta Weights Classification

In this section, our aim is to classify beta weights into their corresponding classes. We follow two scenarios in this task. In the first scenario, we maintain these four distinct classes. In the second scenario, we simplify the classification into two broader categories: Literal (LT) and False (F). Here, the False category combines Literally False, Metaphor, and Scrambled Metaphor classes under a single label "F". The aim of the second scenario is to determine whether the beta weights contain sufficient information for our model to distinguish between stimuli with a true literal meaning and other stimuli.

**Setup** In our classification setup, we followed a structured approach to prepare and train our model. Initially, we standardized the features using StandardScaler to ensure consistent scaling across the dataset. Subsequently, we applied Principal Component Analysis (PCA) to reduce the dimensionality of the data while retaining 85% of the variance, enabling more efficient processing. Each dataset contained a substantial number of NaN values corresponding to ROIs that were not selected. We managed these missing values by either replacing NaNs with zero or dropping affected rows.

Our chosen classifier for this task was Logistic Regression, a widely used algorithm for binary classification tasks that maintains simplicity. To optimize the model's performance, we tuned the hyperparameters of the Logistic Regression classifier using grid search, selecting the best combination of parameters based on cross-validated performance metrics.

**Evaluation** We use confusion matrices to evaluate the performance of our classification model. A confusion matrix is a useful tool that summarizes the performance of a classification algorithm by displaying the counts of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions. Each row of the matrix represents the instances in an actual class, while each column represents the instances in a predicted class. It allows us to identify areas for potential improvement in our classification.

Moreover, to assess the performance of our classification model, we use accuracy. Model accuracy refers to the model's capability of correctly classifying instances in the test dataset. Accuracy is calculated as follows:

$$Accuracy = \frac{Number\ of correct\ predictions}{Total\ number\ of\ predictions} \tag{1}$$

where $NumberofCorrectPredictions$ represents the instances in the test set that the model accurately classified, while the $TotalNumberofPredictions$ denotes the overall number of instances in the test set.

For robust evaluation, we implemented 5-fold nested cross-validation for each subject, ensuring thorough validation and minimizing overfitting. Throughout this process, we maintained the ratio of classes in the test and train sets using stratified cross-validation to preserve the integrity of our model evaluation and ensure balanced performance across different class categories.

### 3.2 Decoding Beta Weights into Sentence Embeddings

In this section, we perform a decoding task to predict sentence stimuli from neural activity, as depicted in Figure 1.

**Setup** There are two primary sub-tasks involved in this task. First, we preprocess the beta weights of brain imaging data. To achieve this, we remove NaN values, standardize them using StandardScaler,

[1] Source code available at `https://github.com/ellietoulabi/CMPUT605`

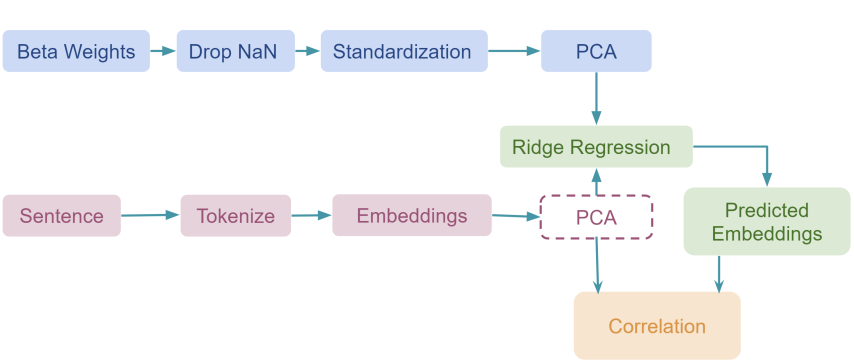



Figure 1: Decoding framework.

and then apply PCA to capture 85% of the data variance, as described in Section 3.1. This ensures that the neural activity data is properly prepared for subsequent analysis.

Secondly, we tokenize our sentences and use Bidirectional Encoder Representations from Transformers (BERT) to generate embeddings for these sentences. These embeddings encapsulate the semantic content of the sentences in a numerical format.

BERT is an advanced language model. It uses a special architecture called Transformer and is trained on a lot of text data using a bidirectional method. Unlike older models that read text in just one direction (like from left to right or right to left), BERT looks at words both before and after a given word at the same time. This helps BERT understand language better by capturing more detailed meanings and connections between words within sentences [15].

In this task, we require a method to convert our sentence stimuli into numerical representations. The representations learned by BERT are utilized for embedding sentences. The [CLS] token vector and middle layer representations of BERT can be used as effective sentence embeddings [16, 17]. In this task, we use middle layer representations of BERT as the embeddings. Figure 2 illustrates the process of using these representations to compute the embedding vector for each sentence.


Figure 2: Generating the embedding vector from BERT's middle layer representation. This process involves averaging over each column of the 5 x 768 matrix to create a 1 x 768 embedding.

Then, we employ PCA to retain either the top 20 most important features or all 768 features of the sentence embeddings, depending on the scenario. PCA helps in reducing the dimensionality of the embeddings while preserving the most relevant information.

Finally, we employ a Ridge Regression for the decoding task. We tune the hyperparameters of the model via grid search to optimize its performance. This comprehensive approach allows us to decode neural representations associated with specific sentence stimuli effectively.

These steps are applied in two scenarios:

- **Scenario 1**: Decoding the embedding for the entire sentence.
- **Scenario 2**: Removing common parts ("Some" and "are") that appear in all sentences to reduce similarity, and then decoding the embedding for the remaining words (denoted as "x y").

**Evaluation** To evaluate the decoding task, we implement nested 5-fold cross-validation, which enhances the robustness and generalizability of our results by dividing the data into multiple train-test splits.

Furthermore, we conduct a permutation test to confirm if the decoding results are statistically significant above chance. This involves shuffling the neuroimaging data and sentence embeddings for the stimuli 100 times to create a null distribution, and comparing our actual results with it. Here, we report the p-value. The p-value is calculated as the proportion of permutations where the test statistic is as extreme as or more extreme than the observed test statistic (Equation 2).

$$p - value = \frac{Number\ of\ permutations\ where\ test\ statistic \geq observed\ test\ statistic}{Total\ number\ of\ permutations} \quad (2)$$

A smaller p-value indicates stronger evidence against the null hypothesis. It means that the observed result is unlikely to have occurred by chance if the null hypothesis were true.

The primary metric used to evaluate the decoding model is the Pearson correlation coefficient between the predicted and actual sentence embeddings. This metric evaluates the similarity of trends between model predictions and representations from the language model across sentences. A higher correlation value indicates a stronger alignment between predicted and actual embeddings, demonstrating the effectiveness of our decoding approach.

Additionally, we report the mean squared error (MSE) between predicted and actual embeddings.

### 3.3 Sentence Embeddings Classification

In this section, we aim to address the question: Do the sentences contain sufficient information about their class for accurate classification? To investigate this, we try to classify the sentence embeddings into their respective classes.

**Setup** In this task, we use the whole-length embeddings (1 x 768) as input data, which are BERT's middle layer representations calculated as explained in Section 3.2. Additionally, in some cases, we use the [CLS] token head of BERT as an embedding. We apply Logistic Regression to classify these embeddings, exploring multiple scenarios:

- **Scenario 1**: We consider all the data across all four classes.
- **Scenario 2**: We use all classes but balance our data labels by selecting a random subset of LT and LF sentences, each with a size of 20, along with all M and SM sentences, also 20 each.
- **Scenario 3**: We classify Metaphorical(M) versus all other classes grouped as Other (O) to assess if metaphorical sentences have embeddings distinct enough from other classes to be differentiated.
- **Scenario 4**: We classify Metaphorical(M) versus a random subset of SM and LF classes grouped as False (F) to examine if the nonliteral meaning of metaphorical sentences makes them differ sufficiently from classes with no literal or nonliteral meaning.
- **Scenario 5**: We focus solely on classifying Metaphors (M) and Scrambled Metaphors (SM). In this scenario, we randomly shuffle the labels and embeddings 100 times to assess the average results based on chance. We also use the cls head of BERT as embedding instead of the middle layer representations.

These scenarios allow us to analyze whether the embeddings of sentences contain enough information to classify them accurately based on their assigned classes.

**Evaluation** In this section, we use accuracy and confusion matrices to assess the classification tasks (refer to Section 3.1 for more details). In all scenarios, we employ a nested 5-fold cross-validation. Additionally, in scenario 5, we utilize a customized 5-fold cross-validation approach.

Some sentences from the Metaphorical (M) and Simile (SM) groups share multiple words, making it challenging to distinguish between these classes. For example:

- **Some_desks_are**_diamonds (SM)

  **Some_desks_are**_junkyards (M)

- **Some_hearts_are**_diseases (SM)

  **Some_hearts_are**_dwellings (M)

  **Some_hearts_are**_ice (M)

  **Some_hearts_are**_zoos (SM)

To mitigate the impact of these shared words on classification, we have designed a cross-validation method that ensures sentences containing three identical words at the beginning are not split between the training and test sets. Instead, these sentences are grouped together and included entirely in either the train or test set, preventing them from being separated across different sets during validation. In the stimuli set, there are 14 groups where two sentences share three words at the beginning and three groups where four sentences share three words at the beginning. Therefore, to design a 5-fold cross-validation, the test set should consist of one of the following:

- 4 groups with 2 sentences each
- 2 groups with 4 sentences each
- 2 groups with 2 sentences each, and 1 group with 4 sentences

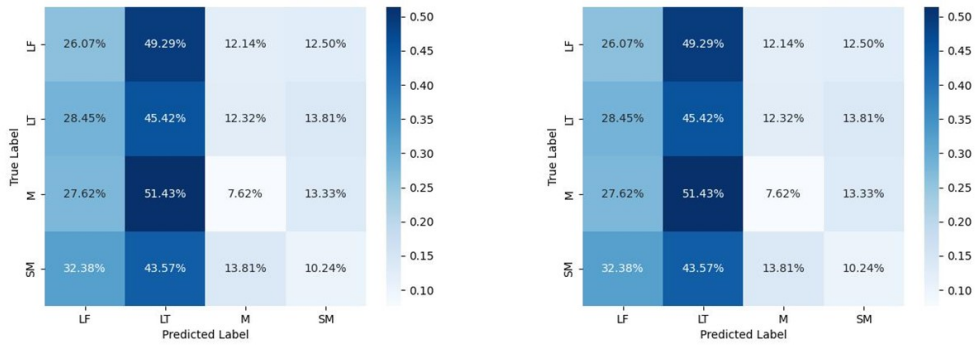
# 4 Results

In this section, we detail the results of our experiments.

## 4.1 Beta Weights Classification

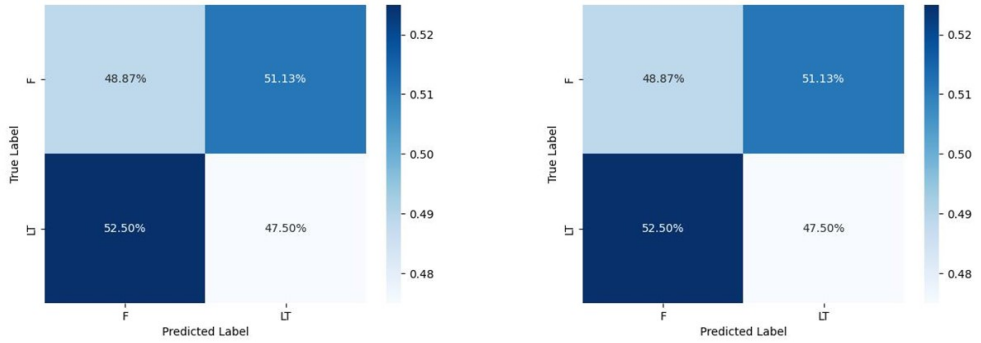
Figure 3 shows the classification results for all four classes of beta weights. In both cases—dropping NaN values and replacing them with zeros—the accuracy is 31.45%. Similarly, the results for both approaches to handling NaN values in the second scenario are identical. Moreover, the accuracy for classifying into two categories is the same (48.18%) for both methods (see Figure 4).

In the first scenario, most instances are classified as LT. In the second scenario, the secondary diagonal elements of the confusion matrix show higher numbers, suggesting that classes are more likely to be misclassified.




(a) (b)

Figure 3: Classification confusion matrices for beta weights considering all four classes. (a) shows results with NaN values replaced by zero, and (b) shows results with NaN values dropped. The matrices show no difference between the two approaches.




(a) (b)

Figure 4: Classification confusion matrices for beta weights considering two classes: F and LT. (a) shows results with NaN values replaced by zero, and (b) shows results with NaN values dropped. The matrices show no difference between the two approaches.

### 4.2 Decoding Beta Weights

Table 2 and 3 present the results of the two scenarios. When applying PCA to the Entire Embeddings approach, there is a decrease in correlation and an increase in MSE. The increase in p-value suggests that the results are more likely due to chance. Similarly, when using embeddings of two words, PCA leads to an increase in correlation and also an increase in MSE. The higher p-value indicates a greater likelihood that these results are by chance. These findings indicate that PCA is not improving the results in this case. This could be due to the limitation of using only 20 features, which may not adequately capture all the information in the data.

By removing common words from the sentences, the correlation increases and the p-value decreases. This suggests that these results are less likely to occur by chance.

Table 2: Results for entire sentence decoding.

| **Method** | **Average Correlation** | **Average MSE** | **Average p-value** |
|---|---|---|---|
| Entire Embeddings | 0.0157 | 0.031 | 0.22 |
| PCA over Embeddings | -0.003 | 0.79 | 0.52 |

Table 3: Results for removing common words decoding (Preserving "x y").

| **Method** | **Average Correlation** | **Average MSE** | **Average p-value** |
|---|---|---|---|
| Entire Embeddings | 0.021 | 0.039 | 0.14 |
| PCA over Embeddings | 0.030 | 0.97 | 0.21 |

### 4.3 Sentence Embeddings Classification

When using all sentence embeddings in an unbalanced setting, the accuracy is 61.87%. Figure 5a illustrates that most LT and LF sentences are classified correctly, but most SM sentences are misclassified as M, and vice versa. When balancing the classes, the accuracy drops to 30%. Similarly, the pattern of misclassifying SM as M and vice versa persists in this scenario too (Figure 5b).

In classifying M versus SM, LT, and LF grouped as "O", the accuracy is 62.5% (see Figure 5c). When classifying M (0) versus SM and LF grouped as "1", the accuracy is 57.5% (Figure 5d). In both cases, the diagonal elements of the confusion matrix display higher values, indicating more correct classifications.

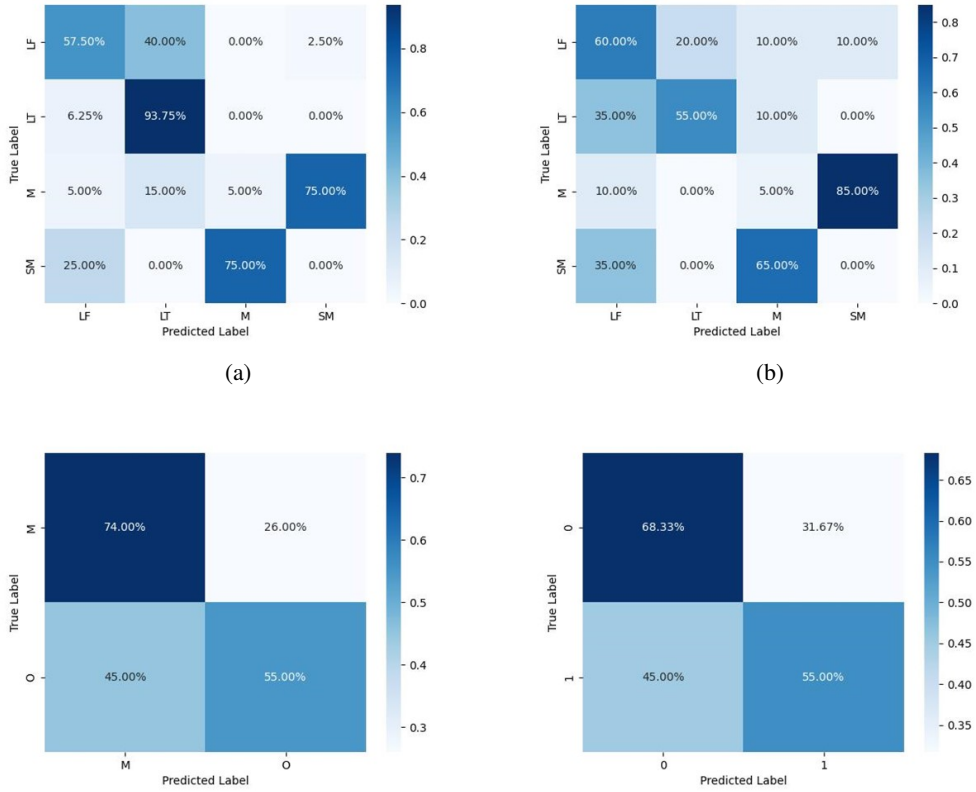



Figure 5: Classification confusion matrices for sentence embeddings. (a) depicts results using all sentence embeddings in an unbalanced setting. (b) displays results for a subsample of LT and LF, along with all M and SM sentence embeddings in a balanced setting. (c) illustrates results for comparisons between M vs. SM, LT, and LF grouped as "O" in a balanced scenario. (d) showcases results for comparisons between M (0) vs. SM and LF grouped as "1" in a balanced setup.

Figure 6 presents the results of classifying M against SM. Initially, we achieved an accuracy of 7.5% on the original data. However, when randomizing the labels and embeddings separately, the accuracies reached to 49.92% and 49.6%, respectively.

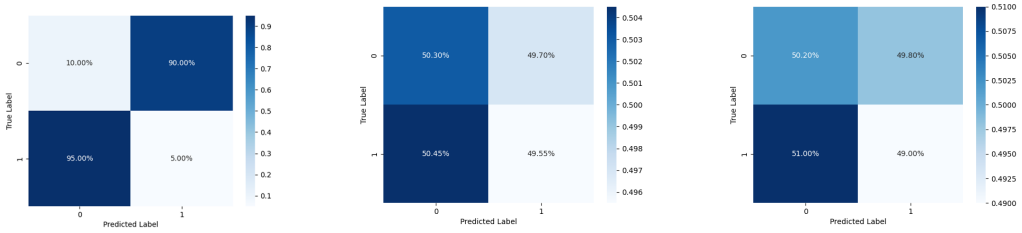
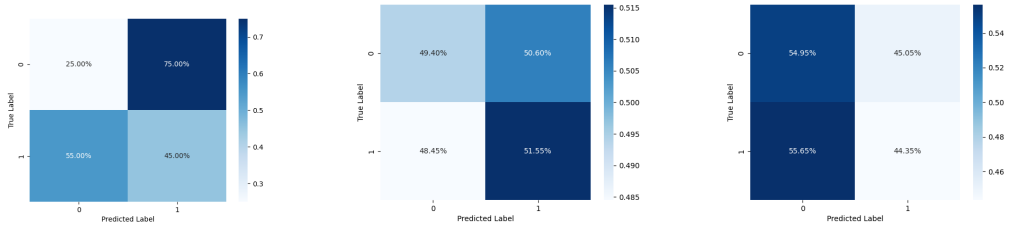



Figure 6: Classification confusion matrices for sentence embeddings of two classes (M and SM). Middle layer representations of BERT are used as embeddings. (a) shows results for the original embeddings. (b) displays results after randomizing labels. (c) illustrates results after randomizing the embeddings.

Figure 7 presents the results of classifying M against SM when we use the CLS head of BERT for embedding. Initially, we achieved an accuracy of 35% on the original data. However, when

randomizing the labels and embeddings separately, the accuracies reached 50.47% and 49.65%, respectively.




(a) (b) (c)

Figure 7: Classification confusion matrices for sentence embeddings of two classes (M and SM). CLS representations of BERT are used as embeddings. (a) shows results for the original embeddings. (b) displays results after randomizing labels. (c) illustrates results after randomizing the embeddings.

## 5 Discussion

In the classification of beta weights into four classes, where NaN values are dropped or replaced with zeros, the majority of classes are mostly classified as either LT or LF. Although the overall accuracy is slightly higher than chance, the classification does not show a strong ability to differentiate between beta weights corresponding to different stimuli classes. Furthermore, in the classification of beta weights into two classes, the classifier is unable to distinguish between beta weights corresponding to LT stimuli with a true literal meaning and other types of stimuli (see Figure 3 and 4.)

As shown in Tables 3 and 2, in the decoding task, when we applied PCA to retain only the 20 most significant features of the embeddings, noticeable changes occurred in our analysis. First, the correlation between variables decreased, and simultaneously, both the MSE and p-value increased while using the entire sentence embeddings. The increase in p-value suggests a higher likelihood that our findings are due to random chance. When we removed the common words "some" and "are" from the sentences, the correlation actually increased with using PCA. However, we observed an increase in both MSE and p-value. Again, this increase in p-value indicates a higher probability that our observed results are coincidental rather than reflecting genuine patterns within the data. These outcomes illustrate that while PCA can be a useful technique for dimensionality reduction, its application in this context could not accurately capture the underlying data structure. A potential explanation could be the insufficiency of the selected number of retained features (20) to capture the essential characteristics of the embeddings.

When common words are removed from sentences, we observe an increase in correlation along with a decrease in the p-value. This suggests that the observed result is less likely to be due to chance compared to using the entire sentence dataset. Removing repeated words improves decoding results. This could be because making sentences less similar helps distinguish them from others more effectively.

At that stage, since the results were not as comprehensive as expected, we looked into the issues affecting the decoding and classification of beta weights. The initial step was to confirm whether the sentences contained sufficient information about their categories for us to classify them. If not, it is unlikely that we would discover meaningful patterns in the fMRI recordings that correspond to these sentences.

To do that, we began by classifying the embeddings of sentences into different groups. When we classified all sentence embeddings (with unbalanced groups), the accuracy was quite high because most of the groups belonged to the LT class. Even if we naively assigned the LT label to all instances, we would still achieve 50% accuracy. Therefore, we balanced the data to get more realistic results. In the balanced dataset, we noticed a significant decrease in accuracy compared to the unbalanced dataset. Upon further analysis, we identified specific patterns of misclassifications within different classes. In many cases, the model predicts M as SM and vice versa. However, it correctly identifies LF as LF and LT as LT most of the time.

Next, we classified M versus SM, LT, and LF grouped together as O (Balanced) to assess M's ability to distinguish itself from other classes. The results showed that out of 12 instances of LT, there were no incorrect predictions, while 2 out of 5 LF instances were misclassified. Additionally, 5 out of 20 M instances and all 3 SM instances were incorrectly predicted. This indicates challenges in accurately distinguishing M from the grouped classes. Furthermore, we observed that the classifier struggled specifically with distinguishing between SM and M.

Afterwards, we attempted to classify M versus SM and LF grouped together as F (Balanced) to assess whether the true non-literal meaning of the M group could differentiate it from other groups. We wanted to understand if the incorrect literal interpretation of M might confuse the classifier. Overall, the model performed well in classifying M correctly. However, there were misclassifications observed in certain cases: 4 out of 12 LF instances were incorrectly predicted, 8 out of 20 M instances were misclassified, and 7 out of 8 SM instances were predicted wrongly. These results suggest that while the classifier was generally effective at distinguishing M, it faces challenges in correctly classifying SM.

Overall, it appears that one of the main challenges is telling apart SM and M. So, we decided to test if we could classify M and SM accurately. In this task, we also addressed the issue of repeated words between M and SM by using our customized cross-validation approach.

In this situation, we noticed that nearly all instances of M were being classified as SM, and vice versa. This surprising pattern led us to randomize the labels and the embeddings. We shuffled labels and embeddings to assess the model's robustness and uncover the true patterns in the data. When we shuffle labels, we mix up the patterns the model learned, which helps us determine if it can still make accurate predictions. Similarly, randomizing embeddings allows us to test the model's performance using different data representations, ensuring that our results are free from biases. We observed that the results were following the same pattern across different classes in randomized cases, but the classifier tended to classify as M slightly more.

We wanted to see how different ways of embedding affect the results. Instead of using the middle layer representations of BERT, we used the CLS head embedding. This change greatly improved classification accuracy, but we still face challenges because the accuracy is still below what we would expect by chance

### 5.1 Ethics

The ethical principles in ML and AI are complex and multifaceted. These principles should be considered during the design, development, and deployment of AI systems [18]. In our project, we used a dataset collected from human participants, emphasizing the importance of ethical considerations throughout. Before participating, all subjects provided informed consent, fully understanding the nature of the experiments. Personal privacy was ensured by anonymizing the data.

It is important to note that limitations in geographic and cultural diversity raise concerns about the generalizability of our findings. Sentences can be interpreted differently across cultures, potentially preventing our results from applying to more diverse populations.

Additionally, the data was gathered in 2016, and since then, there have been advancements in technology, protocols, and methods. This temporal gap challenges the relevance and reliability of our findings.

Furthermore, our evaluation methodology in classification tasks focuses mainly on accuracy and correlation in decoding, potentially overlooking the importance of alternative measures in certain situations. This may introduce evaluation bias, requiring a broader consideration of performance metrics that align with the specific context and goals of the study.

Moreover, to mitigate deployment bias, it is crucial to emphasize that our results should be seen as insights into the process of interpreting metaphors rather than as decision-making rules or directly affecting diagnosis or prognosis factors. Interpretation by qualified individuals is essential, particularly in critical settings, to avoid relying solely on these results for important decisions.

## 6 Future Work

Future investigations could benefit from a whole-brain analysis approach to better understand the neural mechanisms of figurative language comprehension in ASD. Broadening the scope to include additional brain ROIs may uncover novel insights into areas that contribute to metaphor processing.

In parallel, refining sentence embeddings to more accurately reflect metaphorical language presents an avenue for advancing research. The deployment of embeddings that are created to capture the nuances of figurative speech, as mentioned by [19, 20], is anticipated to improve the similarity between computational language models and the neural data from individuals with ASD.

## References


[1] Holly Hodges, Casey Fealko, and Neelkamal Soares. Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation. *Translational Pediatrics*, 9(Suppl 1), 2019.

[2] Joseph S. Alpert. Autism: A spectrum disorder. *The American Journal of Medicine*, 134(6):701–702, 2021.

[3] Ioannis Vogindroukas, Margarita Stankova, Evripidis-Nikolaos Chelas, and Alexandros Proedrou. Language and speech characteristics in autism. *Neuropsychiatric Disease and Treatment*, 18:2367 – 2377, 2022.

[4] Jeffrey D. Rudie Leanna M. Hernandez Susan Y. Bookheimer Natalie L. Colich, Audrey-Ting Wang and Mirella Dapretto. Atypical neural processing of ironic and sincere remarks in children and adolescents with autism spectrum disorders. *Metaphor and Symbol*, 27(1):70–92, 2012. PMID: 24497750.

[5] Patricia Howlin. The acquisition of grammatical morphemes in autistic children: A critique and replication of the findings of bartolucci, pierce, and streiner, 1980. *Journal of Autism and Developmental Disorders*, 14:127–136, 1984.

[6] Margaret Kjelgaard and Helen Tager-Flusberg. An investigation of language impairment in autism: Implications for genetic subgroups. *Language and cognitive processes*, 16:287–308, 05 2001.

[7] Gilbert MacKay and Adrienne Shaw. A comparative study of figurative language in children with autistic spectrum disorders. *Child Language Teaching and Therapy*, 20(1):13 – 32, 2004.

[8] Sam Glucksberg, Patricia Gildea, and Howard B. Bookin. On understanding nonliteral speech: Can people ignore metaphors? *Journal of Verbal Learning and Verbal Behavior*, 21:85–98, 1982.

[9] Brea D. Chouinard and Jacqueline Cummine. All the world's a stage: Evaluation of two stages of metaphor comprehension in people with autism spectrum disorder. *Research in Autism Spectrum Disorders*, 23:107–121, 2016.

[10] Brea Chouinard, Joanne Volden, Ivor Cribben, and Jacqueline Cummine. Neurological evaluation of the selection stage of metaphor comprehension in individuals with and without autism spectrum disorder. *Neuroscience*, 361:19–33, 2017.

[11] James V. Haxby. Multivariate pattern analysis of fmri: The early beginnings. *NeuroImage*, 62(2):852–855, 2012. 20 YEARS OF fMRI.

[12] Nicole Rafidi. Using Machine Learning for Time Series to Elucidate Sentence Processing in the Brain. 7 2019.

[13] Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, 2008.

[14] Kenneth A. Norman, Sean M. Polyn, Greg J. Detre, and James V. Haxby. Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in Cognitive Sciences*, 10(9):424–430, 2006.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[16] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*, 2019.

[17] Hyunjin Choi, Judong Kim, Seongho Joe, and Youngjune Gwon. Evaluation of bert and albert sentence embedding performance on downstream nlp tasks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5482–5487, 2021.

[18] Harini Suresh and John V. Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 2019.

[19] Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online, June 2021. Association for Computational Linguistics.

[20] Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. *ArXiv*, abs/2203.14139, 2022.