# Identifying Contradictory or Implied Meanings in Multilingual Text

**Mohith Lingaraj Mulimani**
*MS in Computer Science*
*University Of Central Florida*

**Nikhil Prakash**
*MS in Computer Science*
*University of Central Florida*

## ABSTRACT

***This project aimed to develop a model capable of detecting entailment, neutrality, and contradiction in multilingual text using TPUs in the NLI domain. To achieve this, two different models were tested: the "bert-base-multilingual-cased" model and the "XLM-RoBERTa large-xnli" model. These models were trained on pre-existing pre-trained models and had some additional preprocessing applied to the input data. The output generated by these models was then fed into a neural network with a dense layer of three neurons, which used the Softmax activation function, the Adam optimizer, and the sparse_categorical_crossentropy loss function.***

***During evaluation, the accuracy, precision, recall, and F1 score were used as evaluation metrics. The results of the experiment showed that the XLM-RoBERTa-large-xnli model performed significantly better than the bert-base-multilingual-cased model. This is because the XLM-RoBERTa model has been trained on a larger corpus compared to the BERT model and is a variant of the BERT model.***

***Overall, this project provides a valuable contribution to the NLI domain by demonstrating the effectiveness of the XLM-RoBERTa-large-xnli model for detecting entailment, neutrality, and contradiction in multilingual text. This approach has the potential to be applied in various fields, including natural language processing and machine learning.***

### Keywords

Entailment, neutrality, contradiction, multilingual text, preprocessing, neural network, NLI, TPUs, multilingual, BERT, XLM-RoBERTa, tokenization, contextualized embedding, natural language inference, dynamic masking.

## 1. INTRODUCTION

Natural Language Processing (NLP) has become an increasingly important field of study in recent years as machine learning models have become capable of performing more complex tasks such as answering questions, generating sentences, and extracting text. NLP is particularly useful in situations where large amounts of textual data need to be analyzed, such as in social media monitoring, customer service, and scientific research.

However, despite the advances made in NLP, one question remains unanswered: can machines establish connections between phrases, or is this still a task that requires human intervention? This is an important question, as the ability of machines to infer connections between phrases has vast implications, ranging from fact-checking to detecting false news and performing text analysis.

In this study, we aim to address the challenge of language inference using a dataset from Kaggle that contains pairs of phrases, consisting of a premise and a hypothesis, from 15 different languages. Language inference is the task of determining whether a hypothesis can be logically inferred from a given premise. To achieve this, tokenizers and embedding models like BERT are typically used.

BERT, a bi-directional transformer, has revolutionized the field of NLP by enabling pre-training on vast amounts of unlabeled textual data to acquire a language representation that can be fine-tuned for specific machine learning applications. By adding just one extra output layer, the pre-trained BERT model can be fine-tuned to provide innovative models for various tasks, including question answering and language inference, without significant architectural changes. BERT has outperformed the state-of-the-art in NLP on several challenging tasks, due to the bidirectional transformer, the novel pre-training tasks of Masked Language Model and Next Structure Prediction, a large amount of data, and Google's powerful computing resources.

In addition to BERT, we also use xlm-roberta-large-xnli as our second model. This model takes xlm-roberta-large and fine-tunes it using a mix of NLI data in 15 languages. Our goal is to build an NLI model that can allocate labels of 0, 1, or 2 to pairings of premises and hypotheses corresponding to entailment, neutrality, and contradiction. To achieve this, we employ pre-trained BERT and xlm-roberta-large-xnli as an embedding generator, followed by classifier models. This approach allows us to extract features useful for downstream tasks, such as classification or language understanding, in any of the 15 languages we are studying.

The ultimate aim of this study is to contribute to the development of more sophisticated NLP models that can handle complex tasks with greater accuracy and efficiency. By comparing the performance of BERT and xlm-roberta-large-xnli on the task of language inference, we hope to shed light on the strengths and weaknesses of these models and provide insights into how they can be further improved to achieve even better results.

## 2. PROBLEM STATEMENT

The ability to accurately determine the relationship between two statements is crucial in many natural language processing tasks. For instance, in machine translation, it is essential to identify whether a target sentence is a translation of a source sentence, which involves understanding the relationship between the two sentences. Similarly, in text summarization, it is important to identify the main idea of a given text, which requires understanding the relationship between the sentences in the text. In addition, in sentiment analysis, understanding the relationship between words and their corresponding sentiments is crucial in determining the overall sentiment of a piece of text.

To tackle these tasks, natural language processing models rely on identifying the relationship between the premise and the hypothesis in a given text. In this study, we aim to explore the various relationships that can exist between two statements, namely Entailment, Neutral, and Contradiction. By providing examples to explain each category, we hope to improve the understanding of these concepts and contribute to the development of more accurate natural language processing models that can identify the relationship between two statements with high precision and recall. Let's examine one example of each of these scenarios based on the following premise:

**Premises:** "She walked out of the room, slamming the door behind her."

- **Hypothesis 1:** She was angry and wanted to express her frustration.
  **Category:** Entailment
  **Explanation:** The premise clearly states that the woman slammed the door behind her, which is often associated with anger or frustration. Therefore, it can be inferred that she was angry and wanted to express her frustration.
- **Hypothesis 2:** She quietly left the room without causing any disturbance.
  **Category:** Contradiction
  **Explanation:** The premise states that the woman slammed the door behind her, which implies that she made a loud noise and caused a disturbance. Therefore, the hypothesis that she quietly left the room contradicts the premise.
- **Hypothesis 3:** She came back into the room a few moments later, smiling and apologizing.
  **Category:** Neutral
  **Explanation:** The hypothesis does not contradict or entail anything from the premise. It simply presents a different scenario that could have happened after the woman walked out of the room.

## 3. RELATED WORK

Unsupervised representation learning has significantly advanced the field of natural language processing and enabled the development of more effective models for various language-related tasks [1,2,3]. This advancement has led to a focus on cross-lingual understanding, which involves training models on one language and applying them to others. This has the potential to facilitate communication and understanding across diverse linguistic and cultural communities, expanding the accessibility and effectiveness of language-related technologies [1,2,3].

Recent research has focused on developing masked language models that can be trained on multiple languages without the need for cross-lingual supervision [4,5]. These models, such as Devlin's mBERT and Conneau's XLM, have demonstrated significant improvements in unsupervised machine translation and sequence pre-training, as well as natural language understanding tasks like cross-lingual natural language inference (XNLI) [5]. Moreover, studies have shown the usefulness of multilingual models such as mBERT [7], and have demonstrated the efficiency of cross-lingual data augmentation for cross-lingual natural language inference [9]. Other studies have shown gains over XLM using cross-lingual multi-task learning [8].

However, all of this work has been done on a smaller scale in terms of training data than the XLM-Roberta technique [10], which focuses on unsupervised learning of cross-lingual representations and their transfer to discriminative tasks. The importance of scaling the amount of data for training language models has also been studied extensively in the literature, with researchers showing that increasing the size of the model as well as the training data can lead to significant performance improvements [6,11]. Inspired by RoBERTa, Conneau et al. [10] demonstrated that mBERT and XLM are undertuned, and that simple improvements in the learning procedure of unsupervised MLM leads to much better performance. XLM-Roberta is trained on cleaned Common Crawls [12], which significantly increases the amount of data for low-resource languages. Similar data has also been shown to be effective for learning high-quality word embeddings in multiple languages.

To build upon this work, the authors of this project used XLM-Roberta as a base method and augmented their data using the method of [5]. For future work, the authors suggest trying cross-lingual data augmentation [9] to further improve the performance of their models.

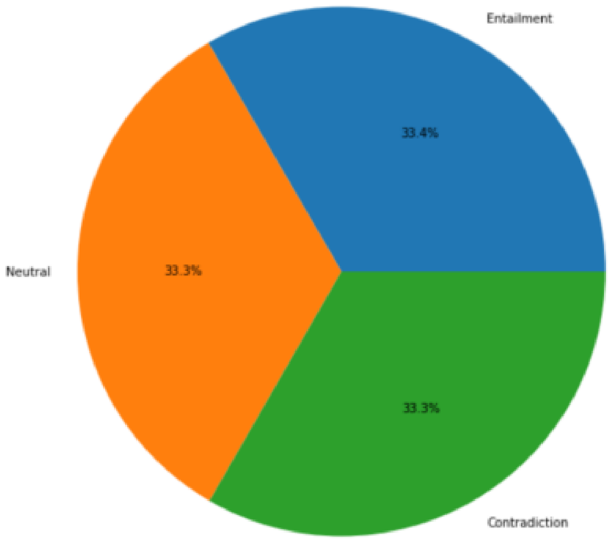
## 4. DATASET and PRE-PROCESSING

The dataset was obtained from Kaggle's website, and it contained various columns such as ID, Premises, Hypothesis, Lang_abv, Language, and label. This dataset was unique because it was multilingual, covering 15 different languages, including English, Hindi, Chinese, and Thai, among others. The label column in the dataset contained three possible values, 0, 1, or 2, which corresponded to the categories of Entailment, Neutral, and Contradiction. The distribution of these labels across the dataset showed that 34.5% of the data was labeled as Entailment, 32% as Neutral, and 33.5% as Contradiction. Overall, there were 12,120 labels in the dataset.

To train and test the machine learning model, we divided the dataset into two sets, namely training and testing sets. The ratio of this division was 80:20, meaning that 80% of the data would be used to train the machine learning model, while the remaining 20% would be used for testing purposes. This division of the dataset allowed us to validate the accuracy of the machine learning model and determine its effectiveness in predicting the categories of Entailment, Neutral, and Contradiction.

During the data preprocessing phase, we performed several steps to clean the data and prepare it for the machine learning model. One of the essential steps was removing some of the features from the dataset. The features we removed, including "id," "language," and "lang_abv," were deemed unnecessary for our analysis, and therefore, we eliminated them. This removal of features aimed to simplify the dataset and reduce any irrelevant noise that could affect the accuracy of our machine learning model's predictions. By performing these data preprocessing steps, we ensured that the data was ready for analysis and that our machine learning model could effectively predict the categories of Entailment, Neutral, and Contradiction.

Below pie chart is the distribution of target class i.e label in the dataset



## 5. TECHNIQUES

The BERT model is a state-of-the-art language processing model that has been pre-trained on a massive corpus of text data in order to learn representations of words and sentences. This pre-training allows the model to be fine-tuned on specific NLP tasks with relatively small amounts of task-specific data. The "bert-base-multilingual-cased" model is a variant of the BERT model that is capable of processing text in multiple languages while retaining case information. It has been trained on a diverse range of languages and is suitable for NLP tasks that involve multilingual text analysis.

The XLM-RoBERTa model, on the other hand, is a modified version of the BERT model that has been developed specifically for cross-lingual applications. This model has been pre-trained on a large corpus of multilingual text data and is capable of encoding text in any of the 100+ languages it has been trained on. The "joeddav/xlm-roberta-large-xnli" model, in particular, has been fine-tuned on the Natural Language Inference (NLI) task using a mix of data from 15 different languages. This fine-tuning allows the model to perform well on a specific NLP task like language inference in the 15 different languages it has been trained on.

The language inference task involves determining the relationship between two statements, which can either be Entailment, Neutral, or Contradiction. The models used in your research have been trained on large amounts of multilingual text data, which allows them to encode text in a way that captures the semantic and syntactic nuances of different languages. By using both the "bert-base-multilingual-cased" and "joeddav/xlm-roberta-large-xnli" models, you can leverage the strengths of each model to improve the accuracy and robustness of your language inference system.

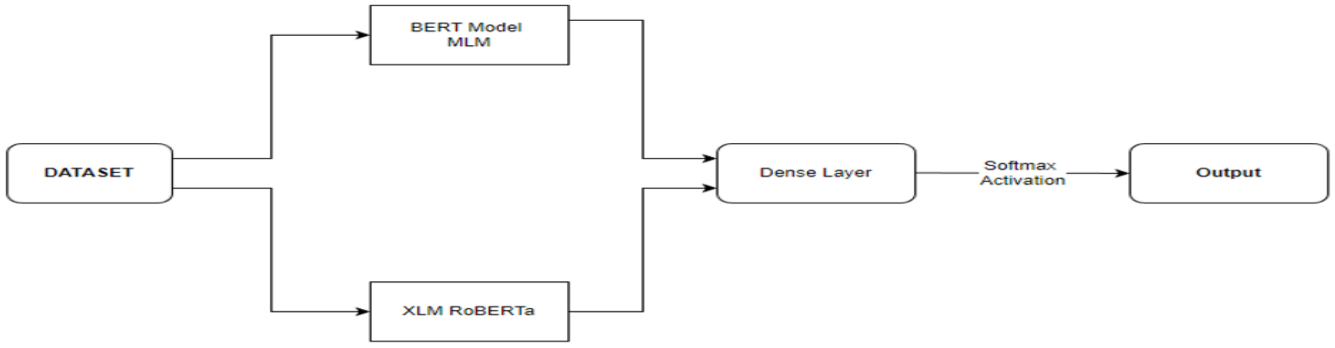
## 5.1 Pipeline

After the dataset has undergone the preprocessing stage, it is split into training, validation, and testing sets. The training and validation sets are used to train and optimize the models, while the testing set is used to evaluate their performance on unseen data.

The preprocessed data is then fed into two state-of-the-art language models, BERT and XLM RoBERTa, which are pre-trained on large amounts of unannotated textual data. The input pairs of premises and hypotheses are tokenized by the models, and additional preprocessing steps such as attention masking and positional embeddings are performed to enhance the models' understanding of the language structure.

The output from the BERT and XLM RoBERTa models are then used as input for a neural network. This neural network consists of a dense layer of 3 neurons, which correspond to the three possible labels: entailment, neutrality, and contradiction. The Softmax activation function is applied to the output layer to generate probability scores for each label.

To train the neural network, the Adam optimizer is used to minimize the sparse_categorical_crossentropy loss function. During training, the model is validated on the validation set to ensure that it is not overfitting to the training data.



Once the model is trained, it is evaluated on the testing set using evaluation metrics such as accuracy, precision, recall, and F1 score. These metrics provide a comprehensive assessment of the model's performance in classifying the relationship between pairs of premises and hypotheses with respect to entailment, neutrality, and contradiction.

In summary, the pipeline involves data preprocessing, training two pre-trained language models (BERT and XLM RoBERTa), generating model outputs, training a neural network, and evaluating the model's performance using evaluation metrics. This approach allows for accurate classification of the relationships between pairs of premises and hypotheses, and can be applied to any of the 15 languages included in the dataset.

## 5.2 BERT Model: bert-base-multilingual-cased

BERT (Bidirectional Encoder Representations from Transformers) is a highly popular pre-trained language model developed by Google. It is capable of understanding text in multiple languages and has been trained on a massive amount of text data from over 100 languages. The model's architecture is based on a transformer neural network, which can capture long-range dependencies in text data by considering the context of each word in both forward and backward directions.

One variant of the BERT model is the bert-base-multilingual-cased, which has been specifically trained on a large corpus of text data from multiple languages while preserving the original capitalization of words. This is particularly useful in languages where capitalization plays a critical role in conveying meaning, as the model can differentiate between words with different capitalizations.

To use the bert-base-multilingual-cased model, we first tokenized the premise and hypothesis using the model's tokenizer. The [SEP] token was added between them to distinguish between the two frames of data. The premise and hypothesis were then converted to tokens and padded with zeros to match the maximum length of the input sequence. To differentiate between the two frames of data, we concatenated the premise and hypothesis with a [CLS] token added at the beginning of the concatenated vector.

To ignore the padded tokens during training, we created a mask that assigned 0 to the padded tokens and 1 to the other tokens. This ensures that the model only focuses on the actual input tokens and ignores the padding. Additionally, we used input_type_ids to differentiate between different sentences, assigning 0 to the premise and 1 to the hypothesis.

After the preprocessing steps, the input_words, input_mask, and input_type_ids were passed as input to the pre-trained bert-base-multilingual-cased model. The model generated a contextualized embedding for the input sentence pair, capturing the meaning and context of the text. This embedding was then used in the subsequent steps to predict the relationship between the premise and hypothesis.

In addition to the bert-base-multilingual-cased model, we also used the XLM-RoBERTa model, which is a variant of the RoBERTa model that has been pre-trained on a large corpus of text data in multiple languages. The XLM-RoBERTa model has achieved state-of-the-art results on a variety of NLP tasks and is particularly useful for cross-lingual applications.

To use the XLM-RoBERTa model, we followed similar preprocessing steps as the bert-base-multilingual-cased model. The main difference is that the XLM-RoBERTa model uses byte-level BPE (Byte Pair Encoding) tokenization, which can better handle the morphological complexity of certain languages. The input sequence was also limited to a maximum length of 512 tokens to ensure computational efficiency.

The output from both the bert-base-multilingual-cased and XLM-RoBERTa models was then fed into a neural network containing a dense layer of 3 neurons, which uses the Softmax activation function, the Adam optimizer, and the sparse_categorical_crossentropy loss function. The evaluation metrics used during testing were accuracy, precision, recall, and F1 score, which were used to classify the relationship between pairs of premises and hypotheses with respect to entailment, neutrality, and contradiction.

### 5.3 XLM-RoBERTa model: XLM-RoBERTa-large-xnli

XLM-RoBERTa-large-xnli is an advanced and refined pre-trained multilingual language model that is built on the success of RoBERTa, which is itself an advanced version of BERT. The main difference between RoBERTa and BERT is that RoBERTa is trained without the next-sentence prediction task and uses larger batch sizes and learning rates during training, resulting in a more powerful and accurate model.

XLM-RoBERTa-large-xnli is trained on a massive amount of data, specifically 2.5TB of filtered Common Crawl data that spans 100 different languages. This means that the model has been trained to understand text in any of these 100 languages and can be fine-tuned for various downstream tasks. During training, the model uses a Masked Language Modeling (MLM) objective, which randomly masks 15% of the words in a sentence, and the model is trained to predict the masked words. This approach allows the model to learn the contextual relationships between words and understand text in a bidirectional manner.

The xlm-roberta-large-xnli model is specifically designed for zero-shot text classification in multiple languages, including those other than English. It is fine-tuned on the XNLI dataset, a multilingual natural language inference dataset, which makes it ideal for use with any language in the XNLI corpus. With its ability to understand and process text in various languages, XLM-RoBERTa-large-xnli can be applied to a wide range of natural language processing tasks, including machine translation, sentiment analysis, and text summarization.

In the implementation of this model, the "xlm-roberta-large-xnli" model is imported from the Hugging Face library, which is a popular NLP library. Next, the Premise and Hypothesis are tokenized, and the [SEP] token is added to differentiate between them. Each token is then converted into its respective ID, and padding zeros are added to match the vector's maximum size. The concatenated vector of Premise and Hypothesis is created by adding a [CLS] token at the beginning, which helps to differentiate the data frame. In this model, input_type_ids are not used, and dynamic masking is applied, where tokens are masked differently at each epoch. Finally, input_words and input_mask are passed as input to the pre-trained model for further processing, which generates a contextualized embedding for the input sentence pair. This contextualized embedding captures the meaning and context of the input text and can be used in downstream tasks such as text classification or language understanding.

## 6. EVALUATION

The two models, BERT-base-multilingual-cased and XLM-RoBERTa-large-xnli, were tested for their performance on a particular task. The first model achieved an accuracy of 65 percent while the second model achieved a much higher accuracy of 89 percent, indicating that the second model performs much better than the first model on the same task.

The evaluation metrics used to test the performance of both models include Accuracy, Precision, Recall, and F1 score, which were calculated on the test data representing 20% of the entire dataset used for the experiment.

| Evaluation Metric | BERT | XLM-RoBERTa |
|---|---|---|
| **Accuracy** | 0.6465 | 0.8874 |
| **Precision** | 0.6486 | 0.8861 |
| **Recall** | 0.6470 | 0.8880 |
| **F1 Score** | 0.6465 | 0.8852 |

Table 1: Experimental Results

Table 1 presents the results of the experiment, showing the accuracy, precision, recall, and F1 score for both models. The XLM-RoBERTa model's higher accuracy and better performance across other evaluation metrics demonstrate that it is a superior model for the given task compared to the BERT model.

## 7. DISCUSSION AND CONCLUSION

The aim of this project is to develop a model that can detect contradiction and entailment in multilingual text using TPUs in the NLI domain. Two models were tested, namely "bert-base-multilingual-cased" and "XLM-RoBERTa large-xnli," with the latter being an improved version of the BERT model. The results of the experiment showed that the XLM-RoBERTa-large-xnli model performed significantly better than the bert-base-multilingual-cased model. This is because the XLM-RoBERTa model has been trained on a larger corpus compared to the BERT model and is a variant of the BERT model.

To further improve the accuracy of the model, different methodologies are being explored. Currently, only one dense layer is being used, and it is necessary to include more dense layers to enhance the accuracy of the model. The dataset consists of 14 languages besides English, and only half of the dataset is in English, while the other half contains the other languages. Therefore, it is essential to increase the dataset for the other languages to ensure that the model is robust enough to handle text in different languages. By doing so, the accuracy of the model can be significantly improved, enabling it to better identify contradictions and entailment in multilingual text, which is crucial in natural language inference.

## 8. REFERENCES

[1] Tomas Mikolov, Quoc V Le, and Ilya Sutskever, “Exploiting similarities among languages for machine translation,” arXiv preprint arXiv:1309.4168, 2013.

[2] Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson, “Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing,” arXiv preprint arXiv:1902.09492, 2019.

[3] Guillaume Lample and Alexis Conneau, “Cross-lingual language model pretraining,” arXiv preprint arXiv:1901.07291, 2019.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pretraining of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[5] Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov, “Xnli: Evaluating cross-lingual sentence representations,” arXiv preprint arXiv:1809.05053, 2018.

[6] Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov, “Emerging cross-lingual structure in pretrained language models,” arXiv preprint arXiv:1911.01464, 2019.

[7] Telmo Pires, Eva Schlinger, and Dan Garrette, “How multilingual is multilingual bert?,” arXiv preprint arXiv:1906.01502, 2019.

[8] Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou, “Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks,” arXiv preprint arXiv:1909.00964, 2019.

[9] Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher, “Xlda: Cross-lingual data augmentation for natural language inference and question answering,” arXiv preprint arXiv:1905.11471, 2019.

[10] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzm´an, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov, “Unsupervised cross-lingual representation learning at scale,” arXiv preprint arXiv:1911.02116, 2019.

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

[12] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzm´an, Armand Joulin, and Edouard Grave, “Ccnet: Extracting high quality monolingual datasets from web crawl data,” arXiv preprint arXiv:1911.00359, 2019