# Contrastive Learning for Universal Zero-Shot NLI with Cross-Lingual Sentence Embeddings

**Md. Kowsher**[1], **Md. Shohanur Islam Sobuj**[2], **Nusrat Jahan Prottasha**[1], **Mohammad Shamsul Arefin**[3], **Yasuhiko Morimoto**[4]

[1]Stevens Institute of Technology, USA
[2]Hajee Mohammad Danesh Science and Technology University, Bangladesh
[3]Chittagong University of Engineering and Technology, Bangladesh
[4]Graduate School of Engineering, Hiroshima University, Japan
{ga.kowsher,shohanursobuj,jahannusratprotta}@gmail.com
sarefin@cuet.ac.bd    morimo@hiroshima-u.ac.jp

## Abstract

Natural Language Inference (NLI) is a crucial task in natural language processing, involving the classification of sentence pairs into entailment, contradiction, or neutral categories. This paper introduces a novel approach to achieve universal zero-shot NLI by employing contrastive learning with cross-lingual sentence embeddings. We utilize a large-scale pretrained multilingual language model trained on NLI data from 15 diverse languages, enabling our approach to achieve zero-shot performance across other unseen languages during the training, including low-resource ones. Our method incorporates a Siamese network-based contrastive learning framework to establish semantic relationships among similar sentences in the 15 languages. By training the zero-shot NLI model using contrastive training on this multilingual data, it effectively captures meaningful semantic relationships. Leveraging the fine-tuned language model's zero-shot learning capabilities, our approach extends the zero-shot capability to additional languages within the multilingual model. Experimental results demonstrate the effectiveness of our approach in achieving universal zero-shot NLI across diverse languages, including those with limited resources. We showcase our method's ability to handle previously unseen low-resource language data within the multilingual model, highlighting its practical applicability and broad language coverage.

## 1 Introduction

Natural Language Processing (NLP) has seen significant advancements in recent years, primarily due to the development of powerful pre-trained languages models like BERT (Devlin et al., 2019a), RoBERTa (Liu et al.), and XLM-RoBERTa (Conneau et al., 2020a). These models have achieved state-of-the-art performance on a wide range of NLP tasks, including Natural Language Inference (NLI) (Bowman et al., 2015; Williams et al., 2018). However, most existing NLI models are limited to the languages they have been explicitly trained on, hindering their applicability across diverse languages and regions. Consequently, there is a growing interest in developing universal zero-shot NLI models capable of generalizing to multiple languages without explicit training data.

Cross-lingual representation learning has emerged as an effective approach to develop models that can understand and process different languages (Ruder et al., 2019). A prominent example is the XLM-RoBERTa model (Conneau et al., 2020a), which leverages a masked language modeling (MLM) objective to learn language-agnostic representations. Despite its effectiveness, XLM-RoBERTa can still benefit from further fine-tuning on specific tasks, such as NLI, to enhance its cross-lingual understanding.

In this paper, we present a novel approach to achieving universal zero-shot Natural Language Inference by leveraging contrastive learning with cross-lingual sentence embeddings depicted in the Figure 1. Our method addresses the challenge of zero-shot NLI, where a model trained on one set of languages can accurately classify sentence pairs in languages it has never seen before. This capability enables the extension of NLI to a vast number of languages without the need for extensive labeled data in each language.

To achieve universal zero-shot NLI, we leverage large-scale pre-trained multilingual language models. Specifically, we utilize an extensively trained multilingual language model, such as XLM-RoBERTa-large, which has been pre-trained on NLI data from 15 diverse languages. This pretraining ensures that the model captures meaningful semantic relationships across different languages.

*Proceedings of the The 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 239–252
December 7, 2023 ©2023 Association for Computational Linguistics

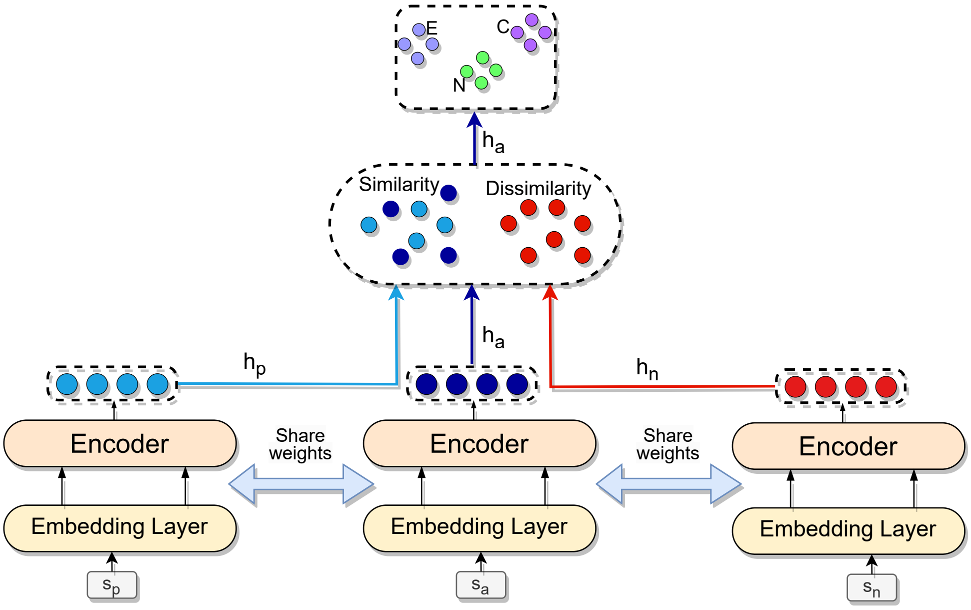

Figure 1: Overview of the proposed methodology for achieving universal zero-shot NLI. The approach incorporates contrastive learning with cross-lingual sentence embeddings, leveraging a large-scale pre-trained multilingual language model trained on NLI data from diverse languages. The Siamese network-based contrastive learning framework establishes semantic relationships among similar sentences, enabling the zero-shot NLI model to capture meaningful semantic representations. By extending the zero-shot capability to additional languages within the multilingual model, the approach achieves universal zero-shot NLI across a broad range of languages, including low-resource ones. In this framework, "a" serves as an anchor, "n" as negative, and "p" as positive in defining the relationships between three categories: entailment (E), neutral (N), or contradiction (C) (for more details, refer to Model Architecture in 5.1).

We exploit the power of contrastive learning by employing a Siamese network-based framework to establish semantic relationships among similar sentences in the 15 languages. Contrastive learning enables the model to learn robust representations that can effectively discriminate between entailment and contradiction.

By training the zero-shot NLI model using contrastive training on this multilingual dataset, we equip the model with the ability to generalize to unseen languages. The fine-tuned language model's zero-shot learning capabilities allow us to extend the zero-shot NLI capability to additional languages within the multilingual model. This approach significantly broadens the language coverage and practical applicability of the NLI model, especially for low-resource languages where labeled data is scarce.

## 2 Related Work

Text classification is a typical task of categorizing texts into groups, including sentiment analysis, question answering, etc. Due to the unstructured nature of the text, extracting useful information from texts can be very time-consuming and inefficient. With the rapidly development of deep learning, neural network methods such as RNN (Hochreiter and Schmidhuber, 1997; Chung et al., 2014) and CNN (Kim, 2014; Zhang et al., 2015) have been widely explored for efficiently encoding the text sequences. However, their capabilities are limited by computational bottlenecks and the problem of long-term dependencies. Recently, large-scale pre-trained language models (PLMs) based on transformers (Vaswani et al., 2017) has emerged as the art of text modeling. Some of these auto-regressive PLMs include GPT (Radford et al., 2018) and XLNet (Yang et al., 2019), auto-encoding PLMs such as BERT (Devlin et al., 2019b), RoBERTa (Liu et al.) and ALBERT (Lan et al., 2019). The stunning performance of PLMs mainly comes from the extensive knowledge in the large scale corpus used for pre-training.

Despite the optimality of the cross-entropy in supervised learning, a large number of studies have revealed the drawbacks of the cross-entropy loss, e.g., vulnerable to noisy labels (Zhang et al., 2018), poor margins (Elsayed et al., 2018) and weak adversarial robustness (Pang et al., 2019). Inspired by the InfoNCE loss (Oord et al., 2018), contrastive learning (Hadsell et al., 2006) has been widely used in unsupervised learning to learn good generic representations for downstream tasks. For example, (He et al., 2020) leverages a momentum encoder to maintain a look-up dictionary for encoding the input examples. (Chen et al., 2020) produces multiple views of the input example using data augmentations as the positive samples, and compare them to the negative samples in the datasets. (Gao et al., 2021) similarly dropouts each sentence twice to generate positive pairs. In the supervised scenario, (Khosla et al., 2020) clusters the training examples by their labels to maximize the similarity of representations of training examples within the same class while minimizing ones between different classes. (Gunel et al., 2021) extends supervised contrastive learning to the natural language domain with pre-trained language models. (Lopez-Martin et al., 2022) studies the network intrusion detection problem using well-designed supervised contrastive loss.

# 3 Background

## 3.1 NLI

Natural Language Inference (NLI) is a task in natural language processing (NLP) where the goal is to determine the relationship between two sentences. Given two input sentences $s_1$ and $s_2$, the task is to classify their relationship as one of three categories: entailment (E), neutral (N), or contradiction (C).

Formally, let $S_1$ and $S_2$ be sets of sentences in two different languages, and let $L = E, N, C$ be the set of possible relationship labels. Given a pair of sentences $(s_1, s_2) \in S_1 \times S_2$, the task is to predict the label $l \in L$ that represents the relationship between the two sentences, i.e., $l = NLI(s_1, s_2)$.

## 3.2 Siamese Networks

Siamese networks are neural network architectures specifically designed for comparing the similarity or dissimilarity between pairs of inputs (Chen and He, 2021). Given two input samples $x_1$ and $x_2$, a Siamese network learns a shared representation for both inputs and measures their similarity based on this shared representation.

Let $f$ denote the shared subnetwork of the Siamese network. The shared subnetwork consists of multiple layers, such as convolutional or recurrent layers, followed by fully connected layers. It aims to extract relevant features from the input samples and map them into a common representation space.

The Siamese network takes two input samples, $x_1$ and $x_2$, and applies the shared subnetwork to each input to obtain the respective representations:

$$h_1 = f(x_1), \quad h_2 = f(x_2)$$

To measure the similarity between $h_1$ and $h_2$, a distance metric is commonly employed, such as Euclidean distance or cosine similarity. For example, cosine similarity can be calculated as:

$$\text{similarity} = \frac{h_1 \cdot h_2}{\|h_1\| \cdot \|h_2\|}$$

During training, Siamese networks utilize a contrastive loss function to encourage similar inputs to have close representations and dissimilar inputs to have distant representations. The contrastive loss penalizes large distances for similar pairs and small distances for dissimilar pairs.

Siamese networks have demonstrated effectiveness in various domains, enabling tasks such as similarity-based classification, retrieval, and clustering. The ability to learn meaningful representations for similarity estimation has made Siamese networks widely applicable in research and practical applications.

## 3.3 Contrastive learning

Let $\mathcal{D} = (x_i, y_i)_{i=1}^{N}$ be a dataset of $N$ samples, where $x_i$ is a sentence and $y_i$ is a label. Let $\phi$ be an embedding function that maps a sentence $x_i$ to a low-dimensional vector representation $\phi(x_i) \in \mathbb{R}^d$, where $d$ is the dimensionality of the embedding space. The goal of contrastive learning is to learn an embedding function $\phi$ such that the similarity between the representation of a sentence $x_i$ and its positive sample $x_j$ is greater than that of its negative samples $x_k$.

Given a pair of sentences $(x_i, x_j)$, the contrastive loss can be defined as follows:

$$
\begin{aligned}
L = -\log \frac{\exp\left(\frac{\mathrm{sim}(x_i,x_j)}{\theta}\right)}{\exp\left(\frac{\mathrm{sim}(x_i,x_k)}{\theta}\right)} \\
+ \sum_{k=1}^{N}[y_k = y_i] \exp\left(\frac{\mathrm{sim}(x_i,x_k)}{\theta}\right) \quad (1) \\
+ \sum_{k=1}^{N}[y_k \neq y_i] \exp\left(\frac{\mathrm{sim}(x_i,x_k)}{\theta}\right)
\end{aligned}
$$

where $\mathrm{sim}(x_i, x_j) = \frac{\phi(x_i)^\top \phi(x_j)}{\|\phi(x_i)\|\|\phi(x_k)\|}$ is the cosine similarity between the embeddings of the sentences $x_i$ and $x_j$, $\theta$ is the temperature parameter that controls the sharpness of the probability distribution over the similarity scores, $[y_k = y_i]$ is the Iverson bracket that takes the value 1 if $y_k = y_i$ and 0 otherwise, and $[y_k \neq y_i]$ is the Iverson bracket that takes the value 1 if $y_k \neq y_i$ and 0 otherwise.

The contrastive loss encourages the model to learn to generate similar embeddings for sentences with the same meaning across different languages, as they will be positively paired during training. This can help enhance the model's cross-lingual understanding and zero-shot learning performance.

## 4 Problem Definition

Let $\mathcal{S}$ denote the set of all sample data, where each sample $s \in \mathcal{S}$ contains multilingual textual data $s^1, s^2, \ldots, s^z \in s$, which are semantically similar. Here, $s_i^z$ represents the $z$-th language data for the $i$-th sample. Each textual data of a language consists of a premise and a hypothesis, separated by a special token, such as [SEP] ($s_{i,p}^z, s_{i,h}^z \in s_i^z$).

The subscripts $p$ and $h$ refer to the hypothesis and premise, respectively.

Now, let $\mathcal{L} = \mathrm{E, N, C}$ be the set of labels for natural language inference (NLI), representing entailment, neutral, and contradiction, respectively. Our objective is to address the task of NLI across multiple languages under the zero-shot learning setting.

Given an input sentence pair $(s_{i,p}^z, s_{i,h}^z)$, the task is to determine their semantic relationship by assigning an NLI label $l \in \mathcal{L}$. We assume limited or no training data is available for some languages, and our goal is to leverage a multilingual pre-trained language model to generalize to unseen languages.

To achieve this, we aim to learn a mapping function $\phi : \mathcal{S} \to \mathbb{R}^d$, where $\phi(s) \in \mathbb{R}^d$ represents the dense vector representation of a sentence $s$ in an embedding space of dimensionality $d$. The embedding function $\phi$ is trained to generate similar embeddings for semantically equivalent sentences across different languages, while producing dissimilar embeddings for sentences with different meanings.

We formulate our NLI model as a multi-task learning problem by simultaneously optimizing two loss functions: the cross-entropy loss and the contrastive loss. The cross-entropy loss is employed to predict the NLI label $l_i$ for a given sentence pair $s_i^z = (s_{i,p}^z, s_{i,h}^z)$. The contrastive loss ensures that cross-lingual sentence embeddings with similar semantics are close together in the embedding space, i.e., for two languages $\alpha, \beta \in z$, $sim(s^\alpha, s^\beta) = \frac{\mathbf{h}^\alpha \cdot \mathbf{h}^\beta}{|\mathbf{h}^\alpha||\mathbf{h}^\beta|} > \tau$, while dissimilar sentence embeddings are far apart, i.e., $sim(s^\alpha, s^\beta) = \frac{\mathbf{h}^\alpha \cdot \mathbf{h}^\beta}{|\mathbf{h}^\alpha||\mathbf{h}^\beta|} < \tau$. Here, $\tau$ represents the similarity threshold.

We optimize both loss functions using stochastic gradient descent with appropriate hyperparameters to train our model for universal zero-shot NLI across multiple languages.

## 5 Methodology

### 5.1 NLI Model Architecture

Let $s_a = s_i^1, s_i^2, \ldots, s_i^z \in \mathcal{S}$ denote the $i$-th sample, considered as the **anchor** batch, which contains $z$ samples from $z$ different languages that are semantically similar. Similarly, we need to find a **negative** batch $s_n$, denoted as $s_n = s_j^1, s_j^2, \ldots, s_j^z \in \mathcal{S}$, where $i \neq j$ and $s_n$ is the farthest from $s_a$ among all samples in $\mathcal{S}$. We employ a clustering approach (Yang et al., 2019) to obtain $s_n$. Initially, we cluster the set $\mathcal{S}$ into $k$ clusters using sentence embedding techniques (Hochreiter and Schmidhuber, 1997). For any text in the $\alpha$-th language in the $i$-th batch, denoted as $s_i^\alpha \in \mathcal{S}$, we determine its corresponding cluster membership, denoted as $\tau_i$. Subsequently, we identify the cluster $\tau_j$ in $s_n$ for the $j$-th batch that is the farthest from the current cluster $\tau_i$, considering it as a non-semantic cluster. From this non-semantic cluster $\tau_j$, we randomly select a sample as $s_n$. During the training phase, we opt for random selection instead of using a deterministic approach. Since we select the $\alpha$-th language for clustering, we refer to it as the clustering priority language. If $C(\cdot)$ represents the trained cluster model, mathematically, we obtain the cluster number of $s_a$ as follows:

$$\begin{aligned} e_a &= T(s_a) \\ \tau_a &= C(e_a) \end{aligned}$$

Here, $T(\cdot)$ is the sentence embedding transformer, and $\tau_a$ is the cluster ID for $s_a$. Now, we need to find the most distant cluster $\tau_n$ by calculating the Euclidean distance between the centers of the two clusters, given by $||c_a - c_n||_2^2$, where $c_a \in \mathbb{R}^d$ is the center of cluster $\tau_a$, and $c_n \in \mathbb{R}^d$ is the center of cluster $\tau_n$.

Next, for every sample in the cluster, we map the farthest distance cluster as $D(\tau_a) = \tau_n$.

Finally, we obtain the most dissimilar batch $s_n$ to $s_a$. To obtain the similar batch $s_p$ (positive), we randomly shuffle $s_a$ to introduce cross-lingual similarity.

The dense vector representation of the $i$-th batch is obtained by passing $s_a$ through the model:

$$h_a = \phi(s_a),$$

where $h_a \in \mathbb{R}^{z \times d}$ represents the hidden state of the $i$-th batch, $z$ is the number of samples (i.e., the total number of languages in $\mathcal{S}$), and $d$ is the embedding space dimensionality.

Using a Siamese network, the hidden states of $s_p$ and $s_n$ are also obtained as follows:

$$\begin{aligned} h_p &= \phi(s_p) \\ h_n &= \phi(s_n) \end{aligned}$$

To measure the similarity between sentences within the $i$-th batch, we define the similarity function $sim(s_{i,a}, s_{i,p})$, which computes the cosine similarity between their embeddings:

$$sim(s_{i,a}, s_{i,p}) = \frac{h_a \cdot h_p}{||h_a|| ||h_p||},$$

The contrastive loss function is used to learn similar embeddings for semantically equivalent sentences across different languages and dissimilar embeddings for semantically non-equivalent sentences across different languages. We combine both the similarity and dissimilarity losses into a single contrastive loss function using the triplet loss, given by:

$$\mathcal{L}c = \sum_{i=1}^{N} \left[ |h_a - h_p|_2^2 - |h_a - h_n| \, 2^2 + \gamma \right]_+ \tag{2}$$

where $\gamma$ is the temperature parameter that controls the smoothness of the similarity function.

The goal of the triplet loss is to encourage the feature vectors for the anchor and positive embeddings to be closer together in the embedding space than the anchor and negative embeddings. The function $[x]_+$ denotes the hinge loss, which penalizes the model if the distance between the anchor and positive embedding is greater than the distance between the anchor and negative embedding by more than a margin $\gamma$.

Here, similar embeddings correspond to semantically equivalent sentences across different languages, and dissimilar embeddings correspond to semantically non-equivalent sentences across different languages.

For the NLI task, the cross-entropy loss is used. Given a sentence pair $(s_p, s_h) \in \mathcal{S}$, the predicted NLI label $p_i$ is obtained as:

$$p_i = G(h_a)$$

where $G(\cdot)$ is a classifier, and $p_i \in \mathbb{R}^{z \times m}$ represents the softmax scores, with $m = 3$ as the number of classes for the NLI labels $\mathcal{L} = \text{E}, \text{N}, \text{C}$.

The cross-entropy loss function is defined as:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{z} \sum_{k=1}^{m} y_{i,k} \log(p_{i,k}),$$

where $y_{i,k}$ is the indicator function, defined as

$$y_{i,k} = \begin{cases} 1, & \text{if the NLI label of the ith batch is } k, \\ 0, & \text{otherwise.} \end{cases}$$

The overall loss function is a combination of the contrastive loss and the cross-entropy loss:

$$L = \mathcal{L}_{\text{C}} + (1 - \lambda)\mathcal{L}_{\text{CE}}$$

where $\lambda$ is a hyperparameter controlling the trade-off between the two losses.

### 5.2 Training for Zero-Shot Classification

The pseudocode for training the NLI model is outlined in Algorithm 1. The algorithm takes as input an NLI multilingual dataset $S$, where $S = S^1, S^2, \ldots, S^z$. Each batch $s_1, s_2, \ldots, s_b$ is randomly sampled from $S$, and the target labels for each batch are denoted as $y_1, y_2, \ldots, y_b$. Additionally, the algorithm utilizes a trained cluster model $C(.)$, a pre-trained masked language model $F(.)$, and a classifier $G(.)$. The objective is to train a universal zero-shot LM model. The training process

**Algorithm 1** Pseudocode for NLI Model Training

```
Require:
  XNLI dataset $S = \{S^1, S^2, \ldots, S^z\}$
  Batch $\{s_1, s_2, \ldots, s_b\} \in S$
  Label for every batch $\{y_1, y_2, \ldots, y_b\} \in Y$
  Trained cluster model $C(.)$
  Pre-trained MLM $F(.)$
  Classifier $G(.)$
  Mapping maximum distance $D(.)$
Ensure: Trained universal zero-shot LM model
  for each epoch do
      for each batch $(s_i, y_i) \in (S, Y)$ do
          $s_a, y_i \leftarrow$ Randomly Shuffle $(s_i, y_i)$
          $s_p \leftarrow$ Randomly Shuffle $s_i$
          $c \leftarrow D(C(s_i^z))$
          $s_n \leftarrow$ Randomly Shuffle $s_c$
          $h_a \leftarrow \phi(s_a)$
          $h_p \leftarrow \phi(s_p)$
          $h_n \leftarrow \phi(s_n)$
          $\hat{y}_i \leftarrow G(h_a)$
          $\mathcal{L}_{CE} \leftarrow L_{CE}(\hat{y}_i, y_i)$
          $\mathcal{L}_C \leftarrow L_C(h_a, h_p, h_n)$
          $\mathcal{L}_{total} \leftarrow \lambda\mathcal{L}_C + \leftarrow (1-\lambda)\mathcal{L}_{CE}$
      end for
      backpropagate and update model parameters using optimizer such as Adam
  end for
```

consists of iterating over each epoch and each batch within an epoch. In each batch, the samples $s_i$ and their corresponding labels $y_i$ are randomly shuffled. Then, a positive batch $s_p$ is created by randomly shuffling $s_i$. The clustering model is used to find the most distant cluster from the current cluster of $s_i$, denoted as $s_c$. A negative batch $s_n$ is created by randomly shuffling the samples in $s_c$. The sentence embeddings $h_a$, $h_p$, and $h_n$ are obtained by passing $s_a$, $s_p$, and $s_n$ through the model function $\phi$. The classifier $G(.)$ predicts the NLI label $\hat{y}i$ for $s_a$. The cross-entropy loss $\mathcal{L}CE$ is computed between $\hat{y}i$ and $y_i$. The contrastive loss $\mathcal{L}C$ is computed using $h_a$, $h_p$, and $h_n$. The total loss $\mathcal{L}_{total}$ is a combination of the contrastive loss and the cross-entropy loss, weighted by the hyperparameter $\lambda$. After computing the loss, the model parameters are updated using an optimizer such as Adam. This process is repeated for each batch in each epoch.

For the training, we use the XNLI dataset (Conneau et al., 2018), which is a multilingual extension of the MNLI dataset. XNLI consists of a few thousand examples from MNLI that have been translated into 15 different languages, including Arabic, Bulgarian, Chinese, English, French, German, Greek, Hindi, Russian, Spanish, Swahili, Thai, Turkish, Urdu, and Vietnamese. The dataset includes three labels: entailment, neutral, and contradiction.

In the hyperparameter configuration, we used a margin of 1.0 for the Triplet loss. The distance metric used was the Euclidean distance, with a 15 batch size. In addition, we used a 8 gradient accumulation step. We used the Adam optimizer during the training procedure, with a decay rate of 0.01. Starting at $2e-6$, the learning rate was linear scheduled.

### 5.3 Fine-Tuning for Zero-Shot Classification

The objective of fine-tuning the NLI model is to enable zero-shot classification, where the model trained on a particular language can work for other unseen languages with similar objectives. The fine-tuning process is similar to NLI training, with a few key differences. In this approach, we do not use a Siamese network architecture. Instead, there is only one forward representation denoted as $h_a$. Additionally, there is no contrastive learning involved.

The fine-tuning process begins by organizing the data in a specific way. We concatenate 60% of the data with its correct label, which is considered as an entailment ($E$) relationship. The remaining 40% of the data is concatenated with another incorrect label, which is considered as a contradiction ($C$) relationship. An example table illustrating the organization of the data is shown in Table 2.

To fine-tune the NLI model, we leveraged rich and resourceful language resources, including English (Maas et al., 2011), (Keung et al., 2020a), Arabic (ElSahar and El-Beltagy, 2015), France (Le et al., 2019), Russian (Fenogenova et al., 2022), Chines (Li et al., 2018). These resources provided diverse and extensive linguistic data for training and enhancing the model's performance. By incorporating data from multiple languages, we aimed to improve the model's generalization capabilities and enable it to handle various languages effectively (Experimental analysis is discussed in the Ablation study 6.4).

## 6 Experiments

We employed two multilingual language models (LM) for our zero-shot learning experiments using the XNLI datasets: XLM-RoBERTa (Conneau

| Dataset \ Model | XLM-RoBERTa | | mDeBERTa-v3 | | mT5 | | mBERT | | mDistilBERT | | XLM-RoBERTa* | | mDeBERTa-v3* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **F1** | **Acc** | **F1** | **Acc** | **F1** | **Acc** | **F1** | **Acc** | **F1** | **Acc** | **F1** | **Acc** | **F1** | **Acc** |
| DKHate | 0.65 | 0.63 | 0.64 | 0.63 | 0.64 | 0.62 | 0.56 | 0.53 | 0.53 | 0.54 | **0.69** | **0.67** | 0.68 | 0.66 |
| + few-shot | 0.68 | 0.66 | 0.67 | 0.66 | 0.67 | 0.67 | 0.60 | 0.58 | 0.55 | 0.54 | **0.71** | **0.70** | **0.71** | 0.69 |
| MARC-ja | 0.78 | 0.79 | 0.80 | 0.80 | 0.80 | 0.81 | 0.73 | 0.70 | 0.53 | 0.53 | **0.81** | **0.82** | **0.81** | **0.82** |
| + few-shot | 0.84 | 0.85 | 0.85 | 0.86 | **0.87** | **0.88** | 0.77 | 0.75 | 0.55 | 0.54 | 0.86 | 0.87 | **0.87** | **0.88** |
| Kor-3i4k | 0.72 | 0.82 | 0.75 | 0.83 | 0.76 | 0.85 | 0.71 | 0.80 | 0.69 | 0.79 | 0.77 | **0.87** | **0.78** | **0.87** |
| + few-shot | 0.75 | 0.86 | 0.77 | 0.87 | 0.78 | 0.88 | 0.73 | 0.82 | 0.72 | 0.81 | **0.79** | 0.88 | **0.79** | **0.89** |
| Id-clickbait | 0.73 | 0.71 | 0.71 | 0.69 | 0.75 | 0.73 | 0.66 | 0.65 | 0.62 | 0.61 | **0.79** | **0.78** | 0.77 | 0.75 |
| + few-shot | 0.76 | 0.74 | 0.75 | 0.72 | 0.80 | 0.80 | 0.69 | 0.69 | 0.67 | 0.68 | **0.83** | **0.83** | 0.81 | 0.81 |
| MCT4 | 0.77 | 0.78 | 0.75 | 0.75 | 0.76 | 0.76 | 0.70 | 0.68 | 0.68 | 0.67 | **0.83** | **0.83** | 0.80 | 0.80 |
| + few-shot | 0.83 | 0.83 | 0.83 | 0.83 | 0.83 | 0.83 | 0.78 | 0.78 | 0.76 | 0.76 | **0.87** | **0.87** | 0.86 | 0.86 |
| MCT7 | 0.74 | 0.75 | 0.75 | 0.75 | 0.76 | 0.76 | 0.72 | 0.71 | 0.68 | 0.67 | **0.79** | 0.78 | **0.79** | **0.79** |
| + few-shot | 0.80 | 0.79 | 0.80 | 0.80 | 0.81 | 0.81 | 0.76 | 0.75 | 0.74 | 0.74 | **0.83** | **0.83** | **0.83** | **0.83** |
| ToLD-br | 0.58 | 0.59 | 0.59 | 0.59 | 0.60 | 0.60 | 0.55 | 0.55 | 0.52 | 0.53 | **0.63** | **0.63** | **0.63** | **0.63** |
| + few-shot | 0.63 | 0.63 | 0.66 | 0.65 | 0.67 | 0.67 | 0.59 | 0.60 | 0.57 | 0.57 | 0.69 | 0.70 | **0.70** | **0.71** |

Table 1: Performance comparison of various multilingual models on unseen and low-resource NLI datasets in both zero-shot and few-shot settings in terms of accuracy, the higher the better. The models with an asterisk (*) denote our proposed universal zero-shot models. The **best results** are highlighted in **bold** and the second best results are highlighted with underline.

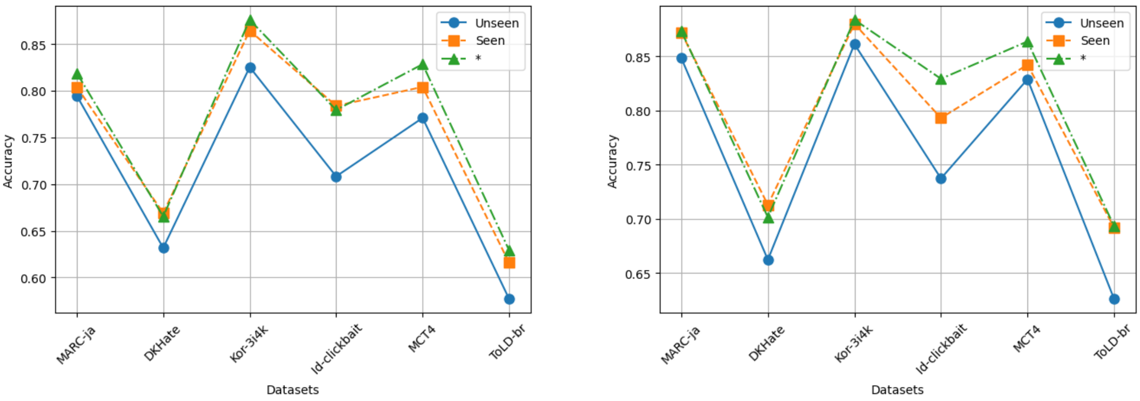

Figure 2: Accuracy comparison of various NLI models in both zero-shot (Left Figure ) and few-shot (Right Figure) settings across different low-resource datasets. The performances of the unseen multilingual XLM-RoBERTa, seen XLM-RoBERTa, and our proposed XLM-RoBERTa* are depicted. In this context, seen alludes to the language data that has been employed in training the zero-shot model, while unseen pertains to data that hasn't been incorporated into the zero-shot training process.

| **Text** | **Label** | **Relationship** |
|---|---|---|
| You are capable of achieving great things | This is an example of positive text | Entailment |
| You are capable of achieving great things | This is an example of negative text | Contradiction |

Table 2: Illustration of text-label relationships for two example sentences, showcasing entailment and contradiction.

et al., 2020b) and mDeBERTa-v3 (He et al., 2023), as outlined in training sections 5.2 and 5.3. In our universal behavior experiments, both models were tested on languages not seen during the zero-shot training phase. Furthermore, we benchmarked our universal zero-shot models against several other prominent multilingual models—mT5 (Xue et al., 2021), mBERT (Devlin et al., 2019b), mDistillBERT (Sanh et al., 2020), XLM-RoBERTa (Conneau et al., 2020b), and mDeBERTa-v3 (He et al., 2023) in a zero-shot setting to gauge their performance. Additionally, we've provided a detailed comparison between our universal zero-shot models and the trained baseline results in Appendix A.3.

## 6.1 Dataset

We used couple of low-resource datasets to conduct the experiment such as MARC-ja (Keung et al., 2020b), DKHate (Sigurbergsson and Derczynski, 2020), kor_3i4k (Cho et al., 2018), id_clickbait (William and Sari, 2020), BanglaMCT (Sobuj et al., 2021), ToLD-Br (Leite et al., 2020). The dataset description is described in the Appendix A.2.

## 6.2 Experimental Results

Based on the presented results in Table 1, our universal zero-shot models, XLM-RoBERTa* and mDeBERTa-v3*, consistently outperformed other multilingual models across various unseen and low-resource datasets. Specifically, in the zero-shot setting, our models achieved the highest accuracy on datasets such as DKHate, MARC-ja, Kor-3i4k, and Id-clickbait. The trend was further emphasized in the few-shot learning scenario, where our models maintained their lead. For instance, on the Id-clickbait dataset, XLM-RoBERTa* achieved an F1 score of 0.83 and an accuracy of 0.83, noticeably surpassing other models. While traditional multilingual models such as mT5 and mBERT demonstrated competitive performance in some scenarios, they did not consistently match the prowess of our proposed models. These results underscore the effectiveness of our approach in handling low-resource languages, emphasizing its potential for broader linguistic applications in the realm of Natural Language Inference.

In Figure 2, we observe a comparative analysis of model accuracy across various unseen and low-resource datasets. Notably, for the zero-shot setting, our proposed XLM-RoBERTa* consistently outperformed the unseen multilingual XLM-RoBERTa and closely matched or even exceeded the performance of the seen version on datasets such as MARC-ja, Kor-3i4k, and MCT4. This trend continues into the few-shot scenario, where our model's accuracy remains competitive, particularly outshining both unseen and seen mDeBERTa-v3 on datasets like Id-clickbait and MCT4. The parity, or in some instances superiority, of our universal zero-shot model compared to the seen model accentuates the potency of our approach, demonstrating its capability to generalize well even to languages it hasn't been explicitly trained on, a crucial trait for practical NLI tasks across diverse linguistic landscapes. More experiment has been described in the Appendix A.3

### 6.3 Ablation Study

### 6.4 Effect of Fine-Tuning on Cross-Lingual

After training a universal zero-shot NLI model, we conducted fine-tuning experiments on specific tasks to assess their impact on cross-lingual sentiment analysis. We utilized a multilingual sentiment analysis dataset (Tyqiangz, 2023) for our evaluation. Initially, we fine-tuned the model on sentiment prediction using datasets in English (En), German (De), Spanish (Es), and French (Fr). Subsequently, we evaluated the model's performance on sentiment analysis tasks in Japanese (Ja), Chinese (Zh), Arabic (Ar), Hindi (Hi), Indonesian (In), Italian (It), and Portuguese (Pt). The results presented in Table 3 demonstrate that fine-tuning for specific tasks in one language significantly enhances sentiment analysis performance across various languages, as measured by Accuracy, Precision, and F1-score metrics.

| Method | Before Fine-tuning | | | After Fine-tuning | | |
|---|---|---|---|---|---|---|
| Language | Acc | Pre | F1 | Acc | Pre | F1 |
| English (En) | 0.51 | 0.53 | 0.52 | 0.54 | 0.55 | 0.55 |
| German (De) | 0.52 | 0.54 | 0.53 | 0.55 | 0.57 | 0.56 |
| Spanish (Es) | 0.50 | 0.52 | 0.51 | 0.53 | 0.55 | 0.54 |
| French (Fr) | 0.53 | 0.55 | 0.54 | 0.56 | 0.58 | 0.57 |
| Japanese (Ja) | 0.51 | 0.53 | 0.52 | 0.54 | 0.56 | 0.55 |
| Chinese (Zh) | 0.50 | 0.52 | 0.51 | 0.53 | 0.55 | 0.54 |
| Arabic (Ar) | 0.50 | 0.52 | 0.51 | 0.53 | 0.55 | 0.54 |
| Hindi (Hi) | 0.52 | 0.54 | 0.53 | 0.54 | 0.57 | 0.56 |
| Indonesian (In) | 0.51 | 0.53 | 0.52 | 0.54 | 0.56 | 0.55 |
| Italian (It) | 0.53 | 0.55 | 0.54 | 0.55 | 0.58 | 0.57 |
| Portuguese (Pt) | 0.52 | 0.54 | 0.53 | 0.54 | 0.55 | 0.56 |

Table 3: Performance Metrics Before and After Fine-Tuning Across Multiple Languages

## 7 Conclusion

In conclusion, this work presents a novel approach to achieving universal zero-shot Natural Language Inference (NLI) across a wide range of languages, including low-resource ones. By leveraging contrastive learning with cross-lingual sentence embeddings and a large-scale pre-trained multilingual language model, we have demonstrated the effectiveness of our approach in capturing meaningful semantic relationships and achieving high-performance NLI classification.

Through the use of a Siamese network-based contrastive learning framework, our approach establishes semantic connections among similar sentences in 15 diverse languages. By training the zero-shot NLI model on this multilingual data, it acquires the ability to generalize to unseen languages, effectively extending the zero-shot capability to a broader range of languages within the multilingual model.

Our experimental findings across different languages and tasks showcase the generalizability and flexibility of our zero-shot approach. By fine-tuning the zero-shot models on a limited amount of task-specific labeled data, we are able to bridge the performance gap and achieve competitive results.

## References

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758.

Won Ik Cho, Hyeon Seung Lee, Ji Won Yoon, Seok Min Kim, and Nam Soo Kim. 2018. Speech intention understanding in a head-final language: A disambiguation utilizing intonation-dependency. *arXiv preprint arXiv:1811.04231*.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Advances in neural information processing systems*, pages 2253–2261.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Unsupervised cross-lingual representation learning at scale.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. Bert: Pre-training of deep bidirectional transformers for language understanding.

Hady ElSahar and Samhaa R El-Beltagy. 2015. Building large arabic multi-domain resources for sentiment analysis. In *Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II 16*, pages 23–34. Springer.

Gamaleldin F Elsayed, Vaishaal Shankar, Ngai-Man Cheung, Nicolas Papernot, and Alexey Kurakin. 2018. Large margin deep networks for classification. In *Advances in Neural Information Processing Systems*, pages 9155–9166.

Muhammad N. Fakhruzzaman, Saidah Z. Jannah, Ratih A. Ningrum, and Indah Fahmiyah. 2021. Clickbait headline detection in indonesian news sites using multilingual bidirectional encoder representations from transformers (m-bert).

Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Tatiana Shavrina, Anton Emelyanov, Denis Shevelev, Alexandr Kukushkin, Valentin Malykh, and Ekaterina Artemova. 2022. Russian superglue 1.1: Revising the lessons not learned by russian nlp models. *arXiv preprint arXiv:2202.07791*.

Xiaohan Gao, Wei Chen, Jing Guo, and Junzhou Huang. 2021. Clr-bert: Contrastive learning for robust pretraining. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 275–287.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2021. Supervised contrastive learning for pretrained language model fine-tuning.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, 2:1735–1742.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A Smith. 2020a. The multilingual amazon reviews corpus. *arXiv preprint arXiv:2010.02573*.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020b. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Md Kowsher, Abdullah As Sami, Nusrat Jahan Prottasha, Mohammad Shamsul Arefin, Pranab Kumar Dhar, and Takeshi Koshiba. 2022. Bangla-bert: transformer-based efficient model for transfer learning and language understanding. *IEEE Access*, 10:91855–91870.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966, Marseille, France. European Language Resources Association.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2019. Flaubert: Unsupervised language model pre-training for french. *arXiv preprint arXiv:1912.05372*.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Yue Li, Xutao Wang, and Pengjian Xu. 2018. Chinese text classification model based on deep learning. *Future Internet*, 10(11):113.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, and Omer and Levy. Roberta: A robustly optimized bert pretraining approach.

Manuel Lopez-Martin, Antonio Sanchez-Esguevillas, Juan Ignacio Arribas, and Belen Carro. 2022. Supervised contrastive learning over prototype-label embeddings for network intrusion detection. *Information Fusion*, 79:200–228.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Tianyu Pang, Chuanxiong Xu, Hongtao Du, Ningyu Zhang, and Jun Zhu. 2019. Improving adversarial robustness via promoting ensemble diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6440–6449.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, 32.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL:* `https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language_understanding_paper.pdf`.

Sebastian Ruder, Ivan Vulic, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–630.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3498–3508, Marseille, France. European Language Resources Association.

Md. Shohanur Islam Sobuj, Md. Kowsher, and Md. Fahim Shahriar. 2021. Bangla multi class text dataset. `https://www.kaggle.com/datasets/shohanursobuj/banglamct`.

Tyqiangz. 2023. Multilingual sentiments dataset. `https://huggingface.co/datasets/tyqiangz/multilingual-sentiments`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Andika William and Yunita Sari. 2020. CLICK-ID: A novel dataset for Indonesian clickbait headlines. *Data in Brief*, 32:106231.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Hongyi Zhang, Moustapha Cisse, and Yann N Dauphin. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *arXiv preprint arXiv:1805.07836*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

# A Appendix

## A.1 Hardware and Software

We perform our experiments on a double NVIDIA RTX3090 GPU with 24GB memory. We use PyTorch (Paszke et al., 2019) as the deep learning framework and Hugging Face's Transformers library (Wolf et al., 2019) to work with the XLM-RoBERTa-large model. We use the official evaluation scripts provided with the XNLI dataset to compute the evaluation metrics.

## A.2 Dataset

The dataset provided in this paper is described in this section.

### A.2.1 MARC-ja

The Multilingual Amazon Reviews Corpus (MARC), from which the Japanese dataset MARC-ja was built (Keung et al., 2020b), was used to create the JGLUE benchmark (Kurihara et al., 2022). This study focuses on text classification, and to that end, 4- and 5-star ratings were converted to the "positive" class, while 1- and 2-star ratings were assigned to the "negative" class. The dev and test set each contained 5,654 and 5,639 occurrences, compared to 187,528 instances in the training set. The extensive collection of product reviews provided by MARC-ja makes it possible to evaluate NLP models in-depth. The characteristics of the dataset and the accuracy metric used for evaluation help to provide a thorough examination of how well models perform on tasks involving Japanese text classification.

### A.2.2 DKHate

The Danish hate speech dataset, used in this study, is a significant resource that consists of anonymized Twitter data that has been properly annotated for hate speech. The dataset offers a targeted and thorough collection for hate speech detection and was produced by Sigurbergsson and Derczynski for their article titled "Offensive Language and Hate Speech Detection for Danish" (Sigurbergsson and Derczynski, 2020). Each element in the collection contains a tweet and a label designating whether or not it is offensive ("OFF" or "NOT"). It has a training split of 2,960 tweets and a test split of 329 tweets.

### A.2.3 Kor-3i4k

The Korean speaker intentions dataset 3i4K used in this study is an invaluable tool for this purpose (Cho et al., 2018). Along with manually crafted commands and inquiries, it includes commonly used Korean terms from the corpus of the Seoul National University Speech Language Processing Lab. It includes classifications for utterances that depend on intonation as well as fragments, statements, inquiries, and directives. This dataset offers essential information on precisely determining speaker intents given the importance of intonation in languages like Korean. With a training set of 55,134 examples and a test set of 6,121 examples, this domain can effectively train and evaluate models.

### A.2.4 Id-clickbait

The CLICK-ID dataset used in this study is made up of a selection of headlines from Indonesian news sources (William and Sari, 2020). There are two primary components to it: Specifically, a subset of 15,000 annotated sample headlines that have been classified as clickbait or non-clickbait and 46,119 raw article data. Three annotators separately examined each headline during the annotation process, and the majority conclusion was taken as the actual truth. There are 6,290 clickbait headlines and 8,710 non-clickbait headlines in the annotated sample. We only trained and evaluated models on the annotated example for the classification task used in this study.

### A.2.5 BanglaMCT

The BanglaMCT dataset, known as the Bangla Multi Class Text Dataset, is a comprehensive collection of Bengali news tags sourced from various newspapers (Sobuj et al., 2021) (Kowsher et al., 2022). It offers two versions, MCT4 and MCT7. MCT4 consists of four tags, while MCT7 includes seven tags. The dataset contains a total of 287,566 documents for MCT4 and 197,767 documents for MCT7. The dataset is split into a balanced 50/50 ratio for training and testing, making it suitable for text classification tasks in Bengali, particularly for news-related content across different categories.

### A.2.6 ToLD-br

The ToLD-Br dataset is a valuable resource for investigating toxic tweets in Brazilian Portuguese (Leite et al., 2020). The dataset provides thorough coverage of LGBTQ+phobia, Xenophobia, Obscene, Insult, Misogyny, and Racism with contributions from 42 annotators chosen to reflect various populations. The binary version of the dataset was used in this study, to evaluate whether a tweet is toxic or not. There are 21,000 examples total in the dataset, with 16,800 examples in the training set, 2,100 examples in the validation set, and 2,100 examples in the test set. This large dataset helps the construction and testing of models for identifying toxicity in Brazilian Portuguese tweets.

## A.3 Universal Zero-shot vs Trained model

In this section, we present the experimental results of our zero-shot and hence few-shot NLI model compared to previously established datasets and trained models. Typically, models that are specifically trained for a task perform better than zero-shot models. However, our models stood up well when compared to these trained models. We demonstrate the performance of our model across various languages and tasks. In our experimental setup, including the training, validation, and test phases, we closely followed the settings defined in the baseline papers.

| Model | Accuracy | |
|---|---|---|
| | Dev | Test |
| Human | 0.989 | 0.990 |
| Tohoku BERT$_{BASE}$ | 0.958 | 0.957 |
| Tohoku BERT$_{BASE}$ (char) | 0.956 | 0.957 |
| Tohoku BERT$_{LARGE}$ | 0.955 | 0.961 |
| NICT BERT$_{BASE}$ | 0.958 | 0.96 |
| Waseda RoBERTa$_{BASE}$ | 0.962 | 0.962 |
| XLM-RoBERTa$_{BASE}$ | 0.961 | 0.962 |
| XLM-RoBERTa$_{LARGE}$ | 0.964 | 0.965 |
| XLM-RoBERTa* | 0.820 | 0.819 |
| + few shot | 0.896 | 0.873 |
| mDeBERTa-v3* | 0.829 | 0.820 |
| + few shot | 0.882 | 0.878 |

Table 4: JGLUE performance on the DEV/TEST sets of the MARC-ja dataset. The $*$ represents our NLI model for zero-shot classification. The baseline performances are taken from (Kurihara et al., 2022)

Table 4 shows the performance of different models on the DEV and TEST sets of the MARC-ja dataset. The baseline models, such as Tohoku BERT$BASE$, Tohoku BERT$LARGE$, NICT BERT$BASE$, Waseda RoBERTa$BASE$, XLM-RoBERTa$BASE$, and XLM-RoBERTa$LARGE$, are explicitly trained models. Our zero-shot models, XLM-RoBERTa$LARGE*$ and mDeBERTa-v3$base*$, initially exhibit lower accuracy but achieve notable improvement after few-shot training. This demonstrates the potential of our zero-shot approach combined with limited fine-tuning data to bridge the performance gap with explicitly trained models.

Table 5 presents the results from sub-task A in Danish. Existing models, such as Logistic Regression DA, Learned-BiLSTM (10 Epochs) DA, Fast-BiLSTM (100 Epochs) DA, and AUX-Fast-BiLSTM (50 Epochs) DA, are trained models. Our zero-shot models, XLM-RoBERTa$LARGE*$ and mDeBERTa-v3$base*$, achieve competitive performance, and their accuracy further improves after few-shot training.

For the FCI module in the Korean language, Ta-

| Model | Macro F1 |
|---|---|
| Logistic Regression DA | 0.699 |
| Learned-BiLSTM (10 Epochs) DA | 0.658 |
| Fast-BiLSTM (100 Epochs) DA | 0.630 |
| AUX-Fast-BiLSTM (50 Epochs) DA | 0.675 |
| XLM-RoBERTa* | 0.685 |
| + few shot | 0.711 |
| mDeBERTa-v3* | 0.680 |
| + few shot | 0.709 |

Table 5: Results from sub-task A in Danish. The baseline performances are taken from (Sigurbergsson and Derczynski, 2020)

| Models | F1 score | accuracy |
|---|---|---|
| charCNN | 0.7691 | 0.8706 |
| charBiLSTM | 0.7811 | 0.8807 |
| charCNN + charBiLSTM | 0.7700 | 0.8745 |
| charBiLSTM-Att | 0.7977 | 0.8869 |
| charCNN + charBiLSTM-Att | 0.7822 | 0.8746 |
| XLM-RoBERTa* | 0.7741 | 0.8760 |
| + few-shot | 0.7913 | 0.8839 |
| mDeBERTa-v3* | 0.7817 | 0.8722 |
| + few-shot | 0.7989 | 0.8901 |

Table 6: Model Performance for FCI module for the Korean Language. The baseline performances are taken from (Cho et al., 2018)

ble 6 displays the performance comparison of different models. Existing models, including charCNN, charBiLSTM, charCNN + charBiLSTM, and charBiLSTM-Att, are trained models. Our zero-shot models, XLM-RoBERTa$LARGE$* and mDeBERTa-v3$base$*, exhibit comparable performance initially and achieve notable improvement after few-shot training.

In the context of clickbait headline detection in Indonesian news sites (Table 7), the average accuracy of established models like M-BERT, Bi-LSTM, CNN, and XGBoost is provided. Our zero-shot models, XLM-RoBERTa$LARGE$* and mDeBERTa-v3$base$*, demonstrate competitive performance initially and show significant enhancement after few-shot training.

Table 8 presents the results of Bengali multiclass text classification. The models compared include biLSTM, CNN, CNN-biLSTM, DNN, Logistic Regression, and MNB. Our zero-shot models, XLM-RoBERTa$LARGE$* and mDeBERTa-v3$base$*, initially show lower accuracy but achieve notable improvement after few-shot training.

Finally, Table 9 displays the model evaluation for toxic language detection in Brazilian Portuguese social media. Existing methods, such as BoW + AutoML, BR-BERT, M-BERT-BR, M-BERT(transfer), and M-BERT(zero-shot), are compared. Our zero-shot models, XLM-RoBERTa$LARGE$* and mDeBERTa-v3$base$*, exhibit competitive performance initially and demonstrate improvement after few-shot training. Overall, our zero-shot NLI models demonstrate the ability to perform reasonably well without explicit training on the target language. Although their initial performance might be lower compared to explicitly trained models, few-shot training significantly

| Model Name | Average Accuracy |
|---|---|
| M-BERT | 0.9153 |
| Bi-LSTM | 0.8125 |
| CNN | 0.7958 |
| XGBoost | 0.8069 |
| XLM-RoBERTa* | 0.7794 |
| + few-shot | 0.8294 |
| mDeBERTa-v3* | 0.7492 |
| + few-shot | 0.8061 |

Table 7: Performance Comparison of Clickbait Headline Detection in Indonesian News Sites. The baseline performances are taken from (Fakhruzzaman et al., 2021)

| | Model | Accuracy | f1-score |
|---|---|---|---|
| MCT4 | biLSTM | 0.9652 | 0.9653 |
| | CNN | 0.9723 | **0.9723** |
| | CNN-biLSTM | 0.9673 | 0.9673 |
| | DNN | 0.9707 | 0.9708 |
| | Logistic Regression | 0.9586 | 0.9587 |
| | MNB | 0.9357 | 0.9359 |
| | XLM-RoBERTa* | 0.8316 | 0.8290 |
| | + few-shot | 0.8713 | 0.8639 |
| | mDeBERTa-v3* | 0.8012 | 0.8007 |
| | + few-shot | 0.8518 | 0.8600 |
| MCT7 | biLSTM | 0.9236 | 0.9237 |
| | CNN | 0.9204 | 0.9204 |
| | CNN-biLSTM | 0.9115 | 0.9114 |
| | DNN | 0.9289 | 0.9290 |
| | Logistic Regression | 0.9156 | 0.9156 |
| | MNB | 0.8858 | 0.8859 |
| | XLM-RoBERTa* | 0.7418 | 0.7562 |
| | + few-shot | 0.8234 | 0.8221 |
| | mDeBERTa-v3* | 0.7441 | 0.7612 |
| | + few-shot | 0.8309 | 0.8237 |

Table 8: Bengali Multi-Class Text Classification Model Performance. The baseline performances are taken from (Sobuj et al., 2021)

| Methods | Precision | Recall | F1-score |
|---|---|---|---|
| BoW + AutoML | 0.74 | 0.74 | 0.74 |
| BR-BERT | 0.76 | 0.76 | 0.76 |
| M-BERT-BR | 0.75 | 0.75 | 0.75 |
| M-BERT(transfer) | 0.76 | 0.76 | 0.76 |
| M-BERT(zero-shot) | 0.61 | 0.58 | 0.56 |
| XLM-RoBERTa* | 0.64 | 0.63 | 0.62 |
| + few-shot | 0.71 | 0.70 | 0.69 |
| mDeBERTa-v3* | 0.64 | 0.62 | 0.62 |
| + few-shot | 0.72 | 0.71 | 0.70 |

Table 9: Model Evaluation for Toxic Language Detection in Brazilian Portuguese Social Media. The baseline performances are taken from (Leite et al., 2020)