DETECTING CONTRADICTION AND ENTAILMENT IN MULTILINGUAL TEXT

Babak Ebrahimi Soorchaei[⋆] *Aakash Shah*[⋆]

[⋆] University of Central Florida

ABSTRACT

Can machines determine the relationships between sentences, or is that still left to humans? If NLP can be applied between sentences, this could have profound implications for fact-checking, identifying fake news, analyzing text, and much more. If we have two sentences, there are three ways they could be related: one could entail the other, one could contradict the other, or they could be unrelated. Natural Language Inferencing (NLI) is a popular NLP problem that involves determining how pairs of sentences (consisting of a premise and a hypothesis) are related. In this project we tackle a problem from the NLI(Natural Language Inference) domain which is detecting contradiction and entailment in multilingual text using TPUs. We tried XLM-Roberta model which is an improved version of BERT model as our base model and improved the results by using XLNI data augmentation. By using this technique our results improved from 92% to 97% on validation set.

1. INTRODUCTION

Natural language processing (NLP) has grown increasingly over the past few years. Machine learning models tackle question answering, text extraction, sentence generation, and many other complex tasks. But, can machines determine the relationships between sentences, or is that still left to humans? If NLP can be applied between sentences, this could have profound implications for fact-checking, identifying fake news, analyzing text, and much more. In this research we are tackling the problem of language inference for a dataset from kaggle website which consists of pair of sentences(a premise and a hypothesis) from the same language and find the relationship between them. There are samples from 15 languages in this dataset. Normally for these task a tokenizer or embedding model like BERT is used. BERT is a bi-directional transformer for pre-training over a lot of unlabeled textual data to learn a language representation that can be used to fine-tune for specific machine learning tasks. . As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications. While BERT outperformed the NLP state-of-the-art on several challenging tasks, its performance improvement could be attributed to the bidirectional transformer, novel pre-training tasks of Masked Language Model and Next Structure Prediction along with a lot of data and Google's compute power. In this project We chose xlm-roberta-large-xnli which takes xlm-roberta-large and fine-tunes it on a combination of NLI data in 15 languages as our base method. Our task is to create an NLI model that assigns labels of 0, 1, or 2 (corresponding to entailment, neutral, and contradiction) to pairs of premises and hypotheses. Our goal is to use a pre-trained Bert as an embedding generator and on top of that add a classifier model to achieve a reasonable result. Then we tried adding XNLI augmented data to improve our results.

2. DATASET AND PROBLEM STATEMENT

We obtain the Dataset from Kaggle website `https://www.kaggle.com/c/contradictory-my-dear-watson/data`. The Dataset Comprises of columns like ID, Premises, Hypothesis, Langabv, Langauge

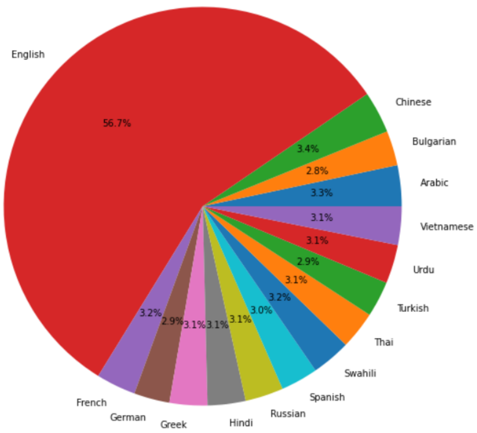


Fig. 1: Language distribution in our dataset

and label. And its for multilingual, it consists of 15 languages like English(majorly), Hindi, Chinese, Thai Et cetera. The Premises and hyposthesis have sentences given, and on Premises base, Hypothesis's sentences will be judged. For example sample of premise is: 'These are issues that we wrestle with in practice groups of law firms, she said.', and sample of hypothesis is: 'Practice groups are not permitted to work on these issues.'. The Judgement or Prediction will be classified into 3 categories i.e. Entailment, Neutral and Contradiction. Let's take a look at an example of each of these cases for the following premise:

- Hypothesis 1 - Just by the look on his face when he came through the door I just knew that he was let down.
  Explanation :We know that this is true based on the information in the premise. So, this pair is related by entailment.

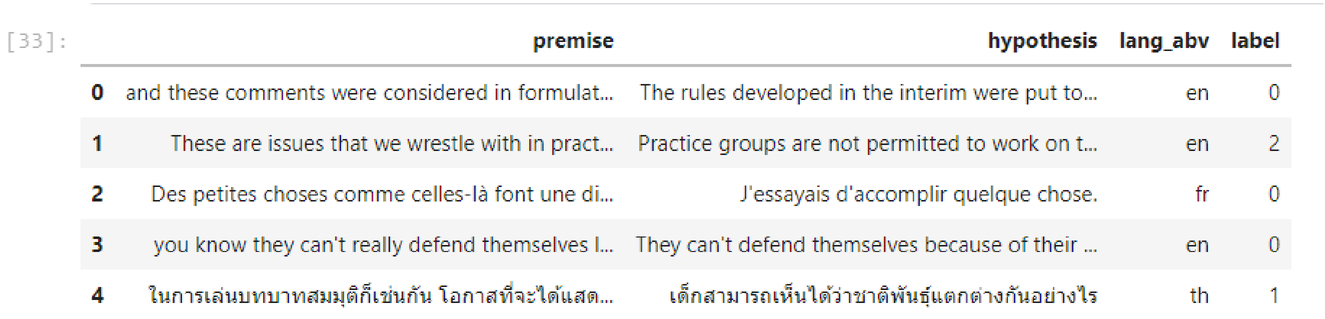
- Hypothesis 2 - He was trying not to make us feel guilty but we knew we had caused him trouble.
  Explanation : This very well might be true, but we can't conclude this based on the information in the premise. So, this relationship is neutral.

- Hypothesis 3 - He was so excited and bursting with joy that he practically knocked the door off it's frame.
  Explanation : We know this isn't true, because it is the complete opposite of what the premise says. So, this pair is related by contradiction.

3. RELATED WORK AND BACKGROUND

Unsupervised representation learning has considerably enhanced the state of the art in natural language understanding, from pretrained word embeddings to pretrained contextualized representations to transformer-based language models. Parallel work on cross-lingual understanding expands these systems to include more languages and the cross-lingual environment, in which a model is learnt in one language and used in another. [1, 2, 3]

Devlin [4] and Conneau[5] recently presented mBERT and XLM, which are masked language models that can be trained on many languages without the need for cross-lingual supervision. On the cross-lingual natural language inference (XNLI) benchmark[5], Lample and Conneau [3] proposed translation language modeling (TLM) as a way to use parallel data to acquire a new state of the art. They also exhibit significant advances in unsupervised machine translation and sequence creation pretraining.[6] demonstrates that monolingual BERT representations are similar across languages, which helps to explain why multilinguality emerges naturally in bottleneck architectures. On sequence labeling challenges, Pires et al. [7] illustrated the usefulness of multilingual models like mBERT.Huang et al. [8] showed gains over XLM using cross-lingual multi-task learning, and Singh et al. [9] demonstrated the efficiency of cross-lingual data augmentation for cross-lingual NLI. However, in comparison to the approach of XLM-Roberta [10] technique, all of this work was done on a much smaller scale in terms of training data. The advantages of scaling language model pretraining by increasing the model's size as well as the training data have been thoroughly researched in the literature. For the monolingual case, researchers showed how large-scale LSTM models can obtain much stronger performance on language modeling benchmarks when trained on billions of tokens. Researchers show in GPT method the importance of scaling



[33]:

	premise	hypothesis	lang_abv	label
0	and these comments were considered in formulat...	The rules developed in the interim were put to...	en	0
1	These are issues that we wrestle with in pract...	Practice groups are not permitted to work on t...	en	2
2	Des petites choses comme celles-là font une di...	J'essayais d'accomplir quelque chose.	fr	0
3	you know they can't really defend themselves l...	They can't defend themselves because of their ...	en	0
4	ในการเล่นบทบาทสมมุติก็เช่นกัน โอกาสที่จะได้แสด...	เด็กสามารถเห็นได้ว่าชาติพันธุ์แตกต่างกันอย่างไร	th	1

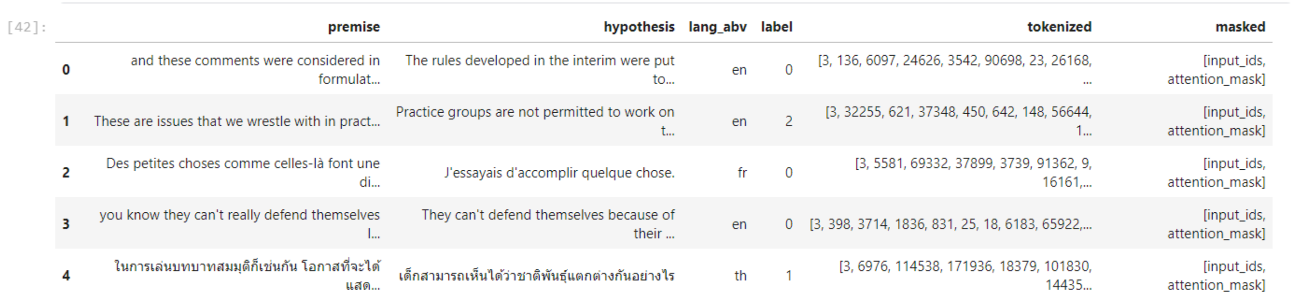
Fig. 2: Sample inputs and classes in our dataset

the amount of data and RoBERTa [11] shows that training BERT longer on more data leads to significant boost in performance. Inspired by RoBERTa, conneau et al.[10] show that mBERT and XLM are undertuned, and that simple improvements in the learning procedure of unsupervised MLM leads to much better performance. XLM-Roberta train on cleaned Common Crawls [12], which increase the amount of data for low-resource languages by two orders of magnitude on average. Similar data has also been shown to be effective for learning high quality word embeddings in multiple languages . Several efforts have trained massively multilingual machine translation models from large parallel corpora. XLM-Roberta, focuses on the unsupervised learning of cross-lingual representations and their transfer to discriminative tasks. In this project we tried XLM-roberta [10] as our base method and then tried to augment our data with method of [5]. For future work we can try cross-lingual data augmentation of [9].

4. MODELS AND TECHNIQUES

We have used model ”joeddav/xlm-roberta-large-xnli”. It is based on XLM-RoBERTa, which is a multilingual version of RoBERTa. RoBERTa builds on BERT and modifies key hyperparameters, removing the next-sentence pretraining objective and training with much larger mini-batches and learning rates.

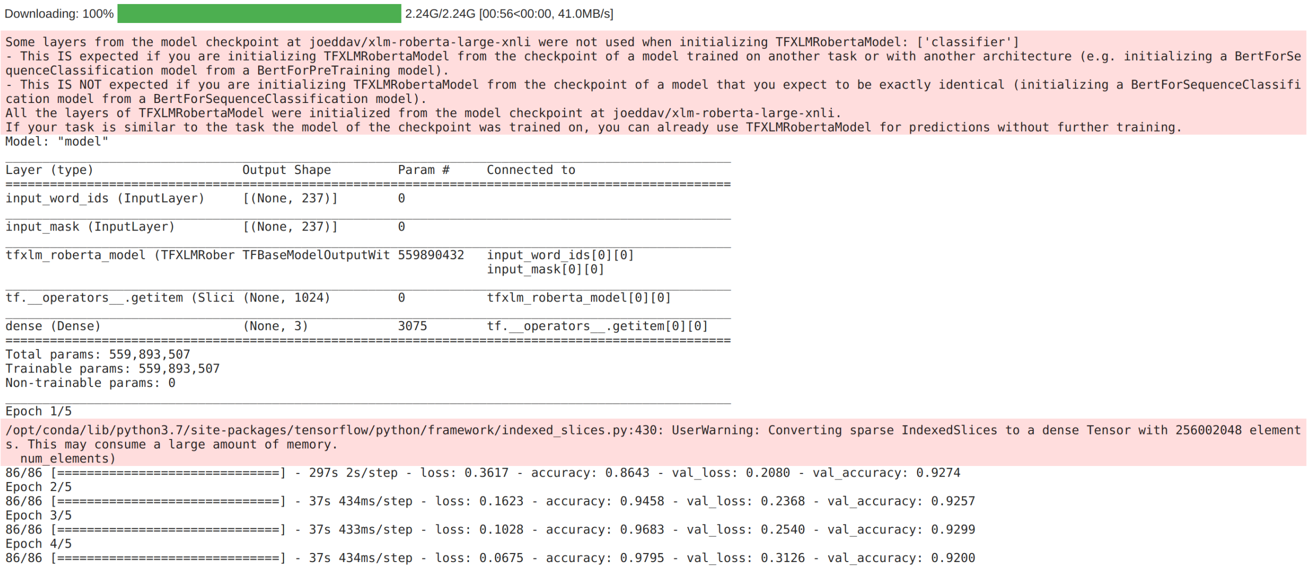
XLM-RoBERTa is pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages.RoBERTa is a transformers model pretrained on a large corpus in a self-supervised fashion. This means it was pretrained on the raw texts only, with no humans labelling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts. More precisely, it was pretrained with the Masked language modeling (MLM) objective. Taking a sentence, the model randomly masks 15% of the words in the input then run the entire masked sentence through the model and has to predict the masked words. This



[42]:

	premise	hypothesis	lang_abv	label	tokenized	masked
0	and these comments were considered in formulat...	The rules developed in the interim were put to...	en	0	[3, 136, 6097, 24626, 3542, 90698, 23, 26168, ...	[input_ids, attention_mask]
1	These are issues that we wrestle with in pract...	Practice groups are not permitted to work on t...	en	2	[3, 32255, 621, 37348, 450, 642, 148, 56644, 1...	[input_ids, attention_mask]
2	Des petites choses comme celles-là font une di...	J'essayais d'accomplir quelque chose.	fr	0	[3, 5581, 69332, 37899, 3739, 91362, 9, 16161,...	[input_ids, attention_mask]
3	you know they can't really defend themselves l...	They can't defend themselves because of their ...	en	0	[3, 398, 3714, 1836, 831, 25, 18, 6183, 65922,...	[input_ids, attention_mask]
4	ในการเล่นบทบาทสมมุติก็เช่นกัน โอกาสที่จะได้แสด...	เด็กสามารถเห็นได้ว่าชาติพันธุ์แตกต่างกันอย่างไร	th	1	[3, 6976, 114538, 171936, 18379, 101830, 14435...	[input_ids, attention_mask]

Fig. 3: sample of inputs and embeddings



```
Downloading: 100% 2.24G/2.24G [00:56<00:00, 41.0MB/s]

Some layers from the model checkpoint at joeddav/xlm-roberta-large-xnli were not used when initializing TFXLMRobertaModel: ['classifier']
- This IS expected if you are initializing TFXLMRobertaModel from the checkpoint of a model trained on another task or with another architecture (e.g. initializing a BertForSe
quenceClassification model from a BertForPreTraining model).
- This IS NOT expected if you are initializing TFXLMRobertaModel from the checkpoint of a model that you expect to be exactly identical (initializing a BertForSequenceClassifi
cation model from a BertForSequenceClassification model).
All the layers of TFXLMRobertaModel were initialized from the model checkpoint at joeddav/xlm-roberta-large-xnli.
If your task is similar to the task the model of the checkpoint was trained on, you can already use TFXLMRobertaModel for predictions without further training.
Model: "model"
__________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
input_word_ids (InputLayer)     [(None, 237)]        0

__________________________________________________________________________________________________
input_mask (InputLayer)         [(None, 237)]        0

__________________________________________________________________________________________________
tfxlm_roberta_model (TFXLMRober TFBaseModelOutputWit 559890432   input_word_ids[0][0]
                                                                 input_mask[0][0]
__________________________________________________________________________________________________
tf.__operators__.getitem (Slici (None, 1024)         0           tfxlm_roberta_model[0][0]
__________________________________________________________________________________________________
dense (Dense)                   (None, 3)            3075        tf.__operators__.getitem[0][0]
==================================================================================================
Total params: 559,893,507
Trainable params: 559,893,507
Non-trainable params: 0
__________________________________________________________________________________________________
Epoch 1/5
/opt/conda/lib/python3.7/site-packages/tensorflow/python/framework/indexed_slices.py:430: UserWarning: Converting sparse IndexedSlices to a dense Tensor with 256002048 element
s. This may consume a large amount of memory.
  num_elements)
86/86 [==============================] - 297s 2s/step - loss: 0.3617 - accuracy: 0.8643 - val_loss: 0.2080 - val_accuracy: 0.9274
Epoch 2/5
86/86 [==============================] - 37s 434ms/step - loss: 0.1623 - accuracy: 0.9458 - val_loss: 0.2368 - val_accuracy: 0.9257
Epoch 3/5
86/86 [==============================] - 37s 433ms/step - loss: 0.1028 - accuracy: 0.9683 - val_loss: 0.2540 - val_accuracy: 0.9299
Epoch 4/5
86/86 [==============================] - 37s 434ms/step - loss: 0.0675 - accuracy: 0.9795 - val_loss: 0.3126 - val_accuracy: 0.9200
```

Fig. 4: Training and validation accuracy after 5 epochs without data augmentation

is different from traditional recurrent neural networks (RNNs) that usually see the words one after the other, or from autoregressive models like GPT which internally mask the future tokens. It allows the model to learn a bidirectional representation of the sentence. This way, the model learns an inner representation of 100 languages that can then be used to extract features useful for downstream tasks: if you have a dataset of labeled sentences for instance, you can train a standard classifier using the features produced by the XLM-RoBERTa model as inputs.

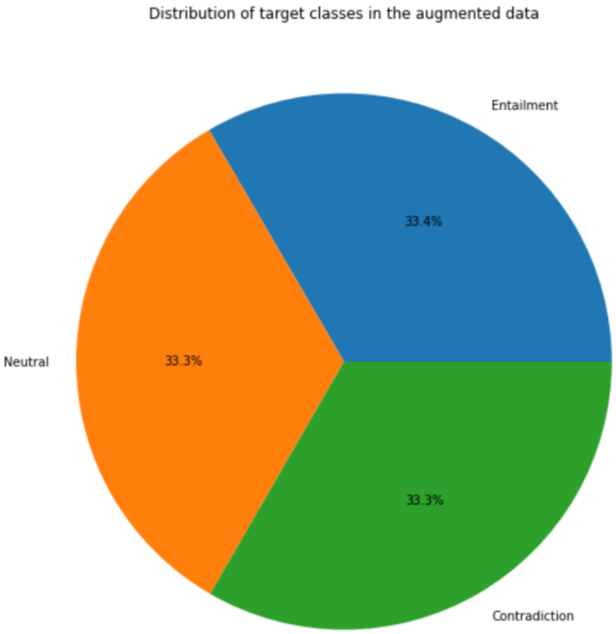
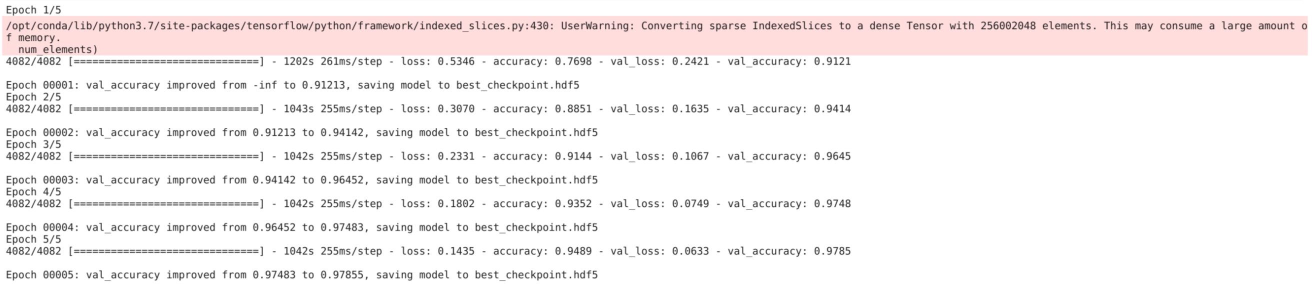


Fig. 5: Distribution of target classes in the augmented data

lm-roberta-large-xnli takes xlm-roberta-large and fine-tunes it on a combination of NLI data in 15 languages. It is intended to be used for zero-shot text classification, such as with the Hugging Face ZeroShotClassificationPipeline. This model is intended to be used for zero-shot text classification, especially in languages other than English. It is fine-tuned on XNLI, which is a multilingual NLI dataset. The model can therefore be used with any of the languages in the XNLI corpus. This model was pre-trained on set of 100 languages. It was then fine-tuned on the task of NLI on the concatenated MNLI train set and the XNLI validation and test sets. Finally, it was trained for one additional epoch on only XNLI data where the translations for the premise and hypothesis are shuffled such that the premise and hypothesis for each example come from the same original English example but the premise and hypothesis are of different languages. By augmenting the XNLI data to our train data in our second experiment we achieved that 4%



```
Epoch 1/5
/opt/conda/lib/python3.7/site-packages/tensorflow/python/framework/indexed_slices.py:430: UserWarning: Converting sparse IndexedSlices to a dense Tensor with 256002048 elements. This may consume a large amount o
f memory.
  num_elements)
4082/4082 [==============================] - 1202s 261ms/step - loss: 0.5346 - accuracy: 0.7698 - val_loss: 0.2421 - val_accuracy: 0.9121

Epoch 00001: val_accuracy improved from -inf to 0.91213, saving model to best_checkpoint.hdf5
Epoch 2/5
4082/4082 [==============================] - 1043s 255ms/step - loss: 0.3070 - accuracy: 0.8851 - val_loss: 0.1635 - val_accuracy: 0.9414

Epoch 00002: val_accuracy improved from 0.91213 to 0.94142, saving model to best_checkpoint.hdf5
Epoch 3/5
4082/4082 [==============================] - 1042s 255ms/step - loss: 0.2331 - accuracy: 0.9144 - val_loss: 0.1067 - val_accuracy: 0.9645

Epoch 00003: val_accuracy improved from 0.94142 to 0.96452, saving model to best_checkpoint.hdf5
Epoch 4/5
4082/4082 [==============================] - 1042s 255ms/step - loss: 0.1802 - accuracy: 0.9352 - val_loss: 0.0749 - val_accuracy: 0.9748

Epoch 00004: val_accuracy improved from 0.96452 to 0.97483, saving model to best_checkpoint.hdf5
Epoch 5/5
4082/4082 [==============================] - 1042s 255ms/step - loss: 0.1435 - accuracy: 0.9489 - val_loss: 0.0633 - val_accuracy: 0.9785

Epoch 00005: val_accuracy improved from 0.97483 to 0.97855, saving model to best_checkpoint.hdf5
```

Fig. 6: Training and validation accuracy after 5 epochs with data augmentation

improvement in our initial results.

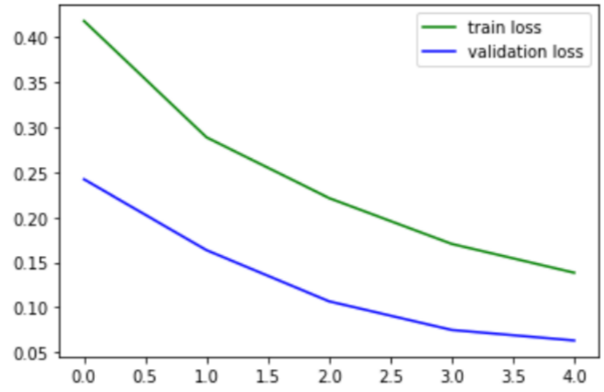


Fig. 7: Train and validation loss during training with the augmented data

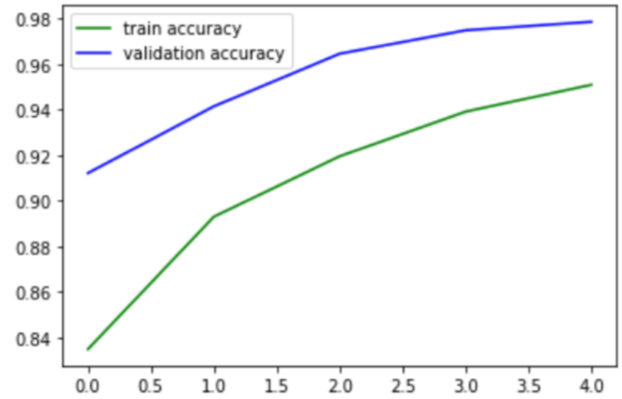


Fig. 8: Train and validation accuracy during training with the augmented data

5. IMPLEMENTATION AND EVALUATION

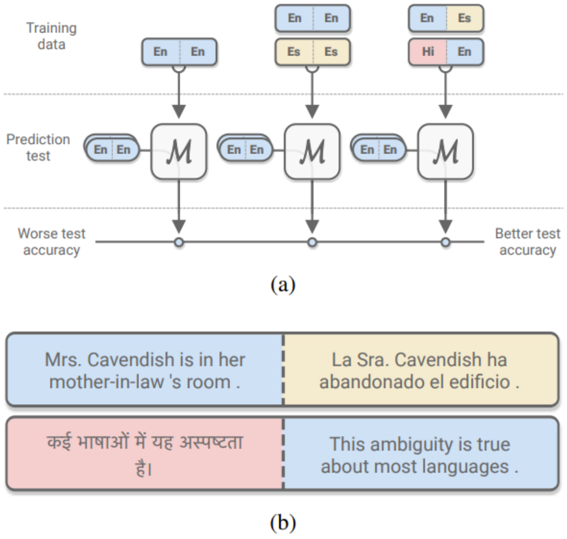
Firstly, We loaded the kaggle dataset with the help of Pandas Library. Then we did some Data Cleaning i.e. remove the columns like Langauge as it is redundant in form of Language abbreviation. Then we imported the BertTokenizer from Transformers and tokenized the sentences. While tokenizing, we added Start of Sentence and Space Separator to make model understand the sentences form. After that, we made embedding vectors for the tokens and applied masked and gave input to the model xlm-roberta-large-xnli for training. After Training for 5 epochs, our Validation Accuracy came out to be about 92 percent. In the next experiment we augmented our data with XNLI dataset and followed the same steps as before. The result of validation set accuracy was more than 97 percent. Figure 6 show the distribution of classes in the train data after data augmentation.

Approach	accuracy on validation set
XLM-RoBERTa without data ugmentation	92%
XLM-RoBERTa with data ugmentation	97%

Table 1: Comparison of accuracy of two implemented approach.

6. DISCUSSION AND CONCLUSION

In this project we tackle a problem from the NLI(Natural Language Inference) domain which is detecting contradiction and entailment in multilingual text using TPUs. We tried XLM-Roberta model which is an improved version of BERT model as our base model and improved the results by using



(a)

Mrs. Cavendish is in her mother-in-law 's room .

La Sra. Cavendish ha abandonado el edificio .

कई भाषाओं में यह अस्पष्टता है।

This ambiguity is true about most languages .

(b)

Fig. 9: Comparing cross-lingual data augmentation. a) compares the standard monolingual approach,a naive multilingual approach that aggregates examples in various languages in a way that each example is solely in one language, and cross-lingual data augmentation (XLDA). For each, prediction is done in a single language. b) two examples of XLDA inputs using the XNLI dataset[9].

data augmentation with XNLI dataset. By using this technique our results improved from 92% to 97% on validation set.

7. FUTURE WORK

Our Future work will be using other approaches for data augmentation like XLDA [9], which cross-lingual data augmentation. We will also try to see if we can boost the performance by using multi-modal separation techniques as we have multi language data in the challenge dataset. The result should be accuracy of prediction which is the percentage of relationships we correctly predict. Also new models like XLM-RoBERTa-XL are on way which can boost the accuracy of classification results in our application.

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: Pre-training 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 Guzmán, 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 Guzmán, Armand Joulin, and Edouard Grave, "Ccnet: Extracting high quality monolingual datasets from web crawl data," *arXiv preprint arXiv:1911.00359*, 2019.