Study of Natural Language Inference in Multilingual Context

Kazi Ababil Azam
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
kaziababilazamtalha@gmail.com

Fardin Anam Aungon
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
fardinanam@gmail.com

ABSTRACT

This project delves into the domain of natural language processing (NLP) with a focus on detecting contradiction and entailment in multilingual text, as explored in the Kaggle competition titled "Contradictory My Dear Watson." The competition provides a challenge of addressing linguistic nuances across different languages, requiring robust models capable of understanding subtle contextual variations.

Our approach involves leveraging state-of-the-art NLP techniques, with a primary emphasis on transformer-based models. The project utilizes pre-trained transformer architectures such as BERT, RoBERTa, and DeBERTa to encode multilingual text representations.

The project's evaluation is based on standard metrics for contradiction and entailment detection, such as accuracy, precision, recall, and F1 score. Experimental results and comparative analyses demonstrate the effectiveness of the proposed approach in addressing the multilingual nature of the competition. In conclusion, this project showcases the significance of advanced NLP techniques in addressing the complex task of contradiction and entailment detection in multilingual text.

KEYWORDS

Natural Language Inference, Multilingual Text, Transformer Models, Kaggle Competition

ACM Reference Format:
Kazi Ababil Azam and Fardin Anam Aungon. 2024. Study of Natural Language Inference in Multilingual Context. In *Proceedings of (CSE472 - Machine Learning Sessional).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Natural Language Inference (NLI) stands as a cornerstone in the realm of Natural Language Processing (NLP), challenging researchers and practitioners to develop models capable of understanding the nuanced relationships between sentences. NLI tasks, such as contradiction and entailment detection, play a pivotal role in advancing comprehension of language semantics and contextual intricacies.

CSE472 - Machine Learning Sessional, March 9, 2024, CSE, BUET

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

In recent years, the advent of transformer-based models has revolutionized the field of NLP. Among these, models like BERT (Bidirectional Encoder Representations from Transformers), RoBERTa (Robustly optimized BERT approach), and DeBERTa have demonstrated exceptional capabilities in capturing contextual information and semantic nuances within sentences. These transformer architectures have become go-to choices for a myriad of NLP applications, showcasing their prowess in tasks ranging from text classification to language understanding.

The Kaggle competition, "Contradictory My Dear Watson," issues a challenge to confront the complexities of multilingual text. The competition signifies the development of models that can detect contradictions and entailments across diverse languages, highlighting the need for cross-lingual understanding.

In this report, we delve into our exploration of advanced NLP techniques applied to the Kaggle competition. Our focus lies in the utilization of transformer-based models, particularly Multilingual BERTs, to address the multilingual nature of the challenge. By incorporating pre-trained models and leveraging transfer learning, we aim to enhance our model's ability to capture language semantics and contextual relationships, ultimately improving performance in contradiction and entailment detection.

This journey takes us through the landscape of NLI models, showcasing the evolution from traditional approaches to the current state-of-the-art transformer architectures. The subsequent sections will detail our methodology, experimentation, and findings, providing insights into the intricate dynamics of multilingual text interpretation and the challenges it entails.

1.1 Problem Definition

Natural language processing (NLP) has grown increasingly elaborate over the past few years. Machine learning models tackle question answering, text extraction, sentence generation, and many other complex tasks. If machines determine the relationships between sentences, that is if NLP can be applied between sentences, this could have profound implications for fact-checking, identifying fake news, analyzing text, and much more.

The Kaggle competition, "Contradictory My Dear Watson," presents a formidable challenge in the form of detecting contradiction and entailment within multilingual text. To comprehend the intricacies of this problem, it's essential to define the fundamental concepts of contradiction and entailment in the context of natural language inference.

In the domain of NLI, the terms "premise" and "hypothesis" play an important role in framing the problem context. The "premise" represents the initial statement or information, while the "hypothesis" is a subsequent statement that is either entailed, contradicted, or is neutral with respect to the premise. In the context of our multilingual contradiction and entailment detection task, these terms gain

significance as we aim to detect the logical relationships between the given premise and hypothesis pairs in diverse languages.

As for the classification of the premise-hypothesis pair in terms of the relationship between them, there are three – entailment, contradiction, and neutral. Each one is explained with an example:

- **Entailment:**
  - **Premise:** "The chef prepared a delectable three-course meal."
  - **Hypothesis:** "A mouthwatering feast was cooked by the chef."

  In this case, the hypothesis is entailed by the premise, as the preparation of a delectable three-course meal implies the creation of a mouthwatering feast.
- **Contradiction:**
  - **Premise:** "The concert was held indoors due to heavy rain."
  - **Hypothesis:** "The outdoor concert was a huge success."

  Here, the hypothesis contradicts the premise, as it suggests the concert's success outdoors, whereas the premise clearly states it was held indoors due to heavy rain.
- **Neutral:**
  - **Premise:** "The student completed the assignment on time."
  - **Hypothesis:** "No information is given about the assignment completion status."

  In this neutral scenario, the hypothesis neither entails nor contradicts the premise; it simply states that there is no provided information about the completion status of the assignment, maintaining a neutral relationship.

In brief, the 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. The train set and test set include languages of fifteen different languages.

2 MATERIALS AND METHODS

The training and test dataset were provided on Kaggle. A starter notebook was also provided as a guide on how to utilize a multilingual BERT model for this task.

2.1 Dataset

The dataset comprises of 12120 pairs of premise and hypothesis with labels marked 0 for entailment, 1 for neutral, and 2 for contradiction. In terms of labels, the dataset seems balanced with around 4000 samples in each class, which is shown in figure 1.

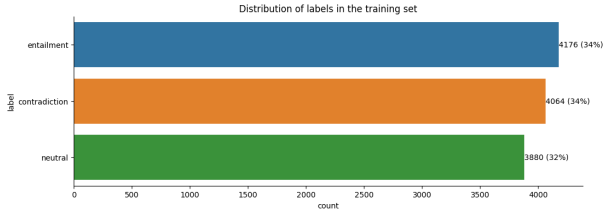

Figure 1: Class Distribution (Training Dataset)

The speciality of this dataset is that it has texts of 15 different languages, which are,

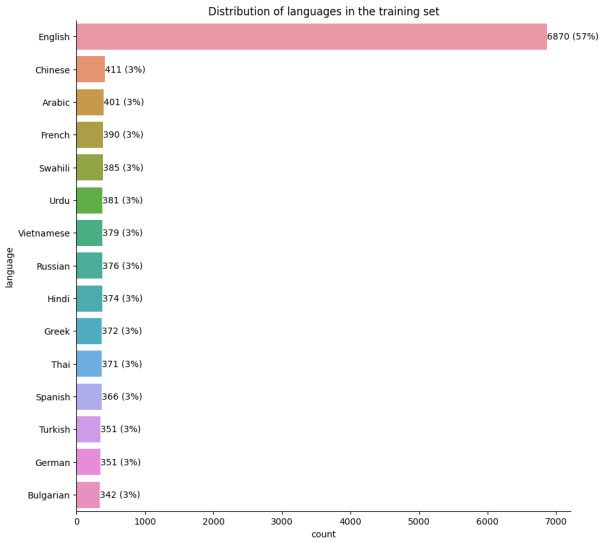

Figure 2: Language Distribution (Training Dataset)

- English
- French
- Thai
- Turkish
- Urdu
- Russian
- Bulgarian
- German
- Arabic
- Chinese
- Hindi
- Swahili
- Vietnamese
- Spanish
- Greek

The number of samples in different languages are unbalanced, as English samples are the vast majority, as can be seen in figure 2.

2.2 Models

2.2.1 ***BERT.*** The BERTClassifier model can be configured with a preprocessor layer, in which case it will automatically apply preprocessing to raw inputs during fit(), predict(), and evaluate(). This is done by default when creating the model with from_preset().

BERT is originally trained on an English corpus. Hence, for tasks involving multiple languages, practitioners often resort to using models specifically designed for multilingual Natural Language Processing (NLP) such as multilingual BERT or XLM-Roberta. In the context of the presented competition, these models prove valuable in handling the diversity of languages present in the dataset.

Here are some models for multi-language NLP available in Keras NLP:

- bert_base_multi
- deberta_v3_base_multi
- distil_bert_base_multi

- xlm_roberta_base_multi
- xlm_roberta_large_multi

The starter notebook utilizes the bert_base_multi.

2.3 Methodology

The method to the solution is braodly two steps:

- **Preprocessing the data:** For NLI, or NLP in general, the data needs to be assigned to tokens and then converted to vectors. This is done using the BERT tokenizer. The tokenizer is used to convert the sentences to tokens and then the tokens are converted to vectors using the BERT model. The vectors are then used as input to the model.
- **Training and finetuining the model:** The model is trained using the training data. The model is then fine-tuned using the validation data.

The different approaches we adopted are as follows:

2.3.1 ***Starter Notebook.*** Our primary approach was to utilize the starter notebook of the competition. It was fairly simple to run and analyze. The preprocessing was done by simply seperating the labels, as the tokenizing is performed by default in the BertClassifier.

The results compiled in the Section 3 show us different scores obtained from the different epochs of training the data.

2.3.2 ***Custom XLMRoBERTa with Wrapper.*** The second approach after a bit of research on multilingual NLI was based off the RoBERTa model, more specifically the XLMRoBERTa model. We used a custom version of the model with a few layers on top of a pretrained XLMRoBERTa.

- An instance of the XLMRobertaModel class initialized with the from_pretrained method using the specified model_name.
- A dropout layer with a dropout rate of 0.2, used for preventing overfitting by randomly setting a fraction of input units to zero during training.
- Classifier Layer:
  - A sequential container for building the classifier part of the neural network, consisting of the following layers:
  - Linear layer (768 input features, 512 output features): Reduces the dimensionality of the input.
  - Layer normalization (512 features): Stabilizes and improves the training of neural networks.
  - Rectified Linear Unit (ReLU) activation function: Introduces non-linearity to the model.
  - Dropout layer with a dropout rate of 0.2: Adds regularization to prevent overfitting.
  - Linear layer (512 input features, 3 output features): The final layer responsible for mapping learned features to the output classes.

It showed better results than the primary approach, so we tried multiple attempts with different parameters and documented the obtained scores.

2.3.3 ***Ensemble Model.*** We tried a different approach to improve our score and position on the leaderboard, which is ensemble learning. We used another pretrained model this time, of multilingual DeBERTa, Decooding-enhanced BERT with disentangled attention. We combine the predictions of the trainer models by stacking them horizontally, resulting in a matrix of predictions where each column corresponds to a model's output for each data point. The combination of predictions can be determined by either selecting the mode (most frequent prediction) or by calculating the mean. The final predictions are obtained by either extracting the mode or rounding the mean along the appropriate axis. This approach allows for combining the strengths of multiple models to improve overall prediction accuracy.

The ensemble model performs as well as the previous approach in terms of score.

2.3.4 ***Pretrained XLMRoBERTa without Wrapper (Best Model).*** After some research and experimentation on the dataset through the previous approaches, we tried using the pretrained model without a classifier on it. The results were far better than we expected, making it the best submission we had on the competition.

We tried the model on the Bangla NLI dataset by BanglaBERT, and it showed better results than the source paper.

3 RESULTS AND DISCUSSION

3.1 Kaggle Competition Dataset

Model	Score
Multilingual Base BERT	0.57343
Multilingual Base BERT	0.61328
Multilingual Base BERT	0.61405
Custom XLMRoBERTa	0.85197
Ensemble DeBERTa	0.86294
Pretrained XLMRoBERTa w/o Wrapper (Best)	**0.8974**

Table 1: Accuracy scores obtained using different approaches to the competition

3.2 Performance of best model on csebuetnlp/xnli_bn

Model	Score
mBERT	67.59
BanglishBERT	70.61
BanglaBERT	72.89
Pretrained XLMRoBERTa w/o Wrapper (Best)	**0.9015**

Table 2: Accuracy scores on csebuetnlp/xnli_bn

4 CONCLUSION

The experimentation led to a lot of different conclusions regarding NLP and NLI, mostly due to the use of different languages in the dataset.

There can be even more improvements in terms of the result, if the languages in which the models do not show satisfactory results are handled seperately. The results on the unincluded Bangla NLI dataset can also be studied, and continued research should improve the quality of research in Bangla NLI and Multilingual Inference in general.