Exploring Transfer Learning Capabilities of Multilingual Language Models

Submitted in fulfillment of the requirements of EE4900: Btech Project

by

Maddineni Bhargava
EE18B112

Under the supervision of

Dr.Gaurav Raina, Dr.Oshin Anand and Dr.Karthika Vijayan

ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY MADRAS

Thesis Certificate

This is to certify that the thesis titled **Exploring Transfer Learning capabilities of Multilingual Language Models**, submitted by **Maddineni Bhargava**, to the Indian Institute of Technology, Madras, for the award of the degree of **Bachelor of Technology**, is a bonafide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Gaurav Raina
Associate Professor
Dept. of Electrical Engineering
IIT-Madras, 600 036

Place: Chennai
Date: May 2022

Abstract

The use of massively multilingual models for natural language processing is becoming increasingly popular in industrial and business applications, particularly in rich multilingual societies. In this paper, we study the capability and extend of transfer learning by multilingual models for text classification and NER in multiple Indian languages for news media archives and various other domains. We train our models with multilingual embedding extractors (mBERT/XLM-R) as the front-end, where training data is made available only in one language. We study the performance characteristics of this classifier model trained in one language when tested in other languages, and observe that the multilingual models showcase transfer learning ability by exploiting the 'inherently parallel' nature of news data. The data that exhibits grossly similar text content across multiple languages, though not in parallel sentences, is termed as inherently parallel in this thesis. Such data exists in scenarios like news articles on same day published in different language editions, customer inquiries/reviews about the same product, social media activity pertaining to same topic, etc. This study reinforces the need of fine-tuning massively multilingual models with in-domain data from a language to express their transfer learning ability in other languages with same domain data. We provide significant evidences to the success of multilingual models for applications with inherently parallel data for easy and automatic maintenance of news media archives.

Contents

Chapter 1

Introduction

There are 7000 major languages in the world, including 120 major languages in India, which creates a divide in terms of the availability of training data and benchmarks for the majority of the world's languages. As a result, the benefits of natural language technology, which have been taken for granted in developing various systems and applications in English and other resource-rich languages are yet to reach many other users. The standard NLP techniques cannot be applied to low-resource languages because they either demand linguistic knowledge which can only be acquired by experts or some knowledge that could only be acquired by native speakers. Furthermore, they might require a large amount of labelled data but manual curation and annotation of large scale resources is time and resource expensive.

The NLP tasks required for text analysis and processing were generally executed by the use of language-specific models and translation/transliteration operations in multilingual scenarios. Such a framework mostly relies on the availability of large amount of labelled data for development of language-specific models and expects to provide solutions to other languages by translation and/or transliteration. This framework of performing NLP tasks has the demerit of accumulation of errors from various stages and the use of multiple models for translation, transliteration, embedding extraction and classification/prediction.

The development of transfer learning provided a more versatile solution to multilingual text processing. The ability to transfer the knowledge of a pre-trained model into a new condition is generally referred to as transfer learning. We transfer either models or sometimes resources in which we can take labelled or unlabeled data created in one language and transfer it to another language in order to quickly develop labelled data in the target language on which to build your models, or we can learn a model in a source language and apply it directly to the target language, thus both transfer of annotations and transfer of models can benefit low resource languages.

Cross-lingual transfer learning is the process of leveraging data and models developed for one language with abundant resources (e.g., English) to address problems in another language with fewer resources. In transfer learning, you can transfer a task developed for one domain to another domain, as well as knowledge from a resource-rich language to a resource-poor language or from one NLP task to another NLP task, so both cross-domain and cross- lingual transfer has benefited NLP systems in general, allowing features and systems developed for one task to benefit other tasks.

In Cross-lingual transfer learning, we train a model for the resource-poor language, but using resources from high-resource languages, we may transfer annotations, generate word or phrase alignments, and use them as bridges between resource-rich and resource-poor languages. We can also project annotations from resource-rich to resource-poor languages. Wheareas in Joint multilingual learning, we train a single model in all languages on a mixed dataset to share data and parameters as much as possible. Thus, the model will learn to generalize a task over all the languages that are involved during the training.

Crosslingual model (XLM) pretraining using causal language modelling, masked language modelling and translation language modelling objectives attempted to improve the performance of multilingual capabilities of NLP. The introduction of Google's bidirectional encoder representations from transformer (BERT) and its variants (RoBERT, ALBERT, DistillBERT, etc.) has revolutionised NLP, which is based on a deep transformer model trained on enormous amount of text data. The BERT learns context of each word from its position based on neighbouring words leading to generation of contextual embeddings, and is pretrained using unlabelled text. The pretrained version of BERT embeddings proves to be beneficial to many downstream tasks when finetuned with a limited amount of task-specific dataset. The success of BERT on English language led to multiple versions of BERT pretrained on a language specific data for many resource-rich languages. But, maintaining a BERT based model for every single language is highly resource expensive

Multilingual language models (MLLM) are extensions of these sophisticated monolingual models. They are trained using huge volumes of unlabelled text from multiple languages and expected to learn embeddings by exploiting similarity between languages like similar vocabulary, genetic relatedness and contact relatedness. Google's multilingual BERT (mBERT) and Facebook AI's crosslingual language model based on RoBERTa (XLM-R) are multilingual language models. The pretrained mBERT has a language representation of 104 languages, while the pretrained XLM-R holds representation of 100 languages and is trained with more data than mBERT. The IndicBERT is a standard multilingual ALBERT model trained on 12 major Indian languages.The MLLMs are almost making language-agnostic text processing a reality. The need of implementation of multilingual NLP is now at an all-time

high due to the reach of conversational AI in rich multilingual societies like India.

1.1 Objective

The primary objective of this thesis is to explore the cross-lingual transfer learning capabilities of pre-trained multilingual language models for various NLP tasks such as Text classification and NER by evaluating their zero-shot performance on multiple Indian languages.

Chapter 2

Literature Review

2.1 Related Works

Cross-lingual transfer learning has made it possible to leverage the capabilities of an NLP model learnt from one or more high resource languages to improve the performance of the model on low resource language in several NLP tasks. Yu-Hsiang Lin et al., 2019 has formulated the task of choosing a set of optimal transfer languages for an NLP task to improve the performance of the model on a given low resource language.

Sparse word representation is one of the earliest attempts to bring in multilingual capabilities to the NLP models. Words from two or more languages are represented in a single semantic space using sparse word embeddings. Gabor Berend et al., 2020 proposed MAMUS(Massively Multilingual Sparse Word Representations) which determines cross-lingually comparable sparse word representations for 27 languages by solving a series of convex optimization problesm. But, it fails to outperform transformer based pretrained multilingual models like mBERT over the 15 languages of the XLNI datasets except in English.

Before the advent of Transformer based language models, a single language-agnostic BILSTM encoder coupled with an auxillary decoder published by Mikel Artetxe et al., 2019 was successful in multilingual

sentence representations for 93 languages.

Yinfei Yang et al., 2019 has introduced three new members in the Universal sentence encoder(USE) (Cer et al., 2018) family of sentence embedding models. Two pre-trained multilingual sentence encoding models based on Transformers and CNN architectures and the third model is an alternative to the transformer model for the retrieval question answering task. These models were successful in representing text from 16 languages in a single semantic space using multi-task tied representations using translation bridge tasks.

Cross-lingual Model Pretraining(XLM)(Guillaume Lample et al.,2019) investingated the impact of cross-lingual model pre-training using Causal Language Modelling(CLM) and Masked language modelling objectives (MLM). It has also introduced Translation Language modelling(TLM) objective which improves cross-lingual language model pretraining.

The introduction of BERT (Bidirectional Encoder representations from Transformers) has revolutionised NLP, resulting in state-of-the-art performance across a wide range of tasks. The procedure comprises training a deep transformer-based model on massive volumes of monolingual data before fine-tuning it with small amounts of task-specific data. The encoder is pre-trained with a masked language modelling goal, and the final result is an encoder that learns great sentence representations. These pre-trained phrase representations improve performance on downstream tasks when fine-tuned on even small amounts of task-specific training data. This formula has been duplicated across languages because to its effectiveness in English NLP, resulting in a plethora of language-specific BERTs. However, only a few languages with the required data and computing resources are able to train such language-specific models.

Multilingual language models (MLLMs) such as mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), XLM-R (Conneau et al., 2020a), and others have grown popular as an alternative. A MLLM is trained using huge volumes of unlabeled data from different languages with the hope that low resource languages might get benefited from high resource languages due to similar vocabulary, genetic relatedness, or contact relatedness. Several such MLLMs have recently been proposed, with differences in architecture (number of layers, parameters, etc.), objective functions used for training (masked language modelling objective, causal language modelling objective, translation language modelling objective, etc.), data used for pretraining (Wikipedia, CommonCrawl, etc.) and number of languages involved during the pre-training. Sumanth Doddapaneni et al., 2021 (A Primer on Pretrained Multilingual Language Models) reviewed the existing literature on MLLMs and proposed ways to buid bigger and better MLLMs using different resources and architectures.

Fangxiaoyu Feng et al., 2020 proposed a 109-language BERT sentence embedding model that is language agnostic, successfully presented an approach to adopt a pre-trained BERT model to a dual encoder model to train the crosslingual embedding space efficiently. Telmo Pires et al., 2019 explored the multilingual capabilities of mBERT. It shows that transfer learning works best for typologically similar languages and mBERT's multilingual representation is able to map learned structures onto new vocabularies but it does not seem to learn systematic transformations of those structures to accommodate a target language with different word order. While Shihie Wu et al., 2020 says that out of 104 languages covered by mBERT, 30% of languages with the least pretraining resources perform worse than using no pretraining model at all. Also, training a monolingual model on low resource lan-

guages does no better. Training on pairs of closely related low resource languages helps but still lags behind mBERT. On the other hand, the highest resource languages (top 10%) are hurt by massively multilingual joint training. While mBERT has access to numerous languages, the resulting model is worse than a monolingual model when sufficient training data exists.

Conneau et al., 2020 shows that XLM-R achieves state-of-the-art performance on cross-lingual classification, sequence labelling and question answering. It outperforms the previous state of the art by 5.1% average accuracy on XNLI, 2.42% average F1-score on Named Entity Recognition, and 9.1% average F1-score on cross-lingual Question Answering.

While models like XLMR and mBERT successfully exhibits multilingual capabilities in NLP tasks like text classification, NER and question answering. They cannot be directly fine-tuned for Natural Language Generation(NLG) downstream tasks like text summarization, machine translation, etc. Yinhan Liu et al., presented mBART, a sequence to sequence denoising auto-encoder pre-trained on large-scale monolingual corpora in many languages using the BART(Lewis et al., 2019) objective and showed that multilingual de-noising pretraining cab improve both supervised and unsupervised machine translation at the sentence and document levels. mT5(Linting Xue et al., 2021) is a massively multilingual text to text transformer model, pretrained on Common Crawl dataset covering 101 languages following a similar recipe as T5 model(Colin Raffel et al., 2020). It has achieved state-of-the-art performance on many cross-lingual NLP and NLG tasks.

2.1.1 mBERT

mBERT stands for multilingual BERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2018). It is essentially just BERT model pre-trained on Wikipedia dataset which consists

104 languages with a shared vocabulary across all languages. BERT model is a stacked representation of Transformer's encoder layer. It uses MLM(Masked Language Modelling) objective function, where 15% of the words in each sequence are replaced with a [MASK] token and the model is trained to predict the original value of the masked token, based on the context provided by the non-masked words in the sequence.

2.1.2 XLM-R

XLM-R (Conneau et al., 2019) is a transformer based multilingual masked language model pre-trained on 2.5TB of CommonCrawl data in 100 languages, which obtains state-of-the-art performance on cross-lingual classification, sequence labelling and question answering. It is pre-trained in a RoBERTa (Liu et al., 2019) fashion using only the MLM objective. Despite handling 100 languages, XLM-R is competitive with monolingual models on a monolingual benchmark. XLM-R achieves an average performance score of 91.5 compared to 90.2, 92.0 and 92.8 of BERT, XLNet and RoBERTa respectively. So while XLM-R outperforms BERT, it is remarkably close to its monolingual counterpart RoBERTa.

2.1.3 IndicBERT

IndicBERT (Divyanshu Kakwani et al.,) is a multilingual ALBERT model. It is pre-trained using the standard masked language model(MLM) objective, exclusively on *IndicCorp* consisting 12 major Indian languages with a monolingual corpus of around 9 billion tokens and subsequently evaluated on a diverse set of tasks. IndicBERT has much fewer parameters than other multilingual models such as mBERT, XLM-R, etc. It also achieves a performance on-par or better than these models on tasks in Indian languages. The 12 languages covered by IndicCorp are: Assamese, Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, Telugu.

Chapter 3

Data

3.1 Text Classification :

3.1.1 News Classification :

The scarcity of task-specific labelled datasets in many languages is a key problem while undertaking NLP studies on Indian languages. We primarily used the IndicNLP News Article Classification Dataset for the Text Classification task. This collection contains news articles in Bengali, Gujarati, Kannada, Malayalam, Malayalam, Marathi, Oriya, Punjabi, Tamil, and Telugu in many areas such as entertainment, sports, business, lifestyle, technology, crime, politics, and so on. A few open source datasets from multiple sources in Hindi, Tamil and Telugu have also been used.

Language	Classes	Number of articles per Class
Telugu	entertainment, business, sports	8K
Tamil	entertainment, politics, sports	3.9K
Malayalam	entertainment, business, sports, technology	1.5K

Kannada	entertainment, lifestyle, sports	10K
Bengali	entertainment, sports	7K
Gujarati	entertainment, business, sports	0.68K
Marathi	entertainment, lifestyle, sports	1.5K
Oriya	entertainment, business, sports,crime	7.5K
Punjabi	entertainment, business, sports, politics	0.78K

Table 3.1: IndicNLP News Classification dataset details

3.1.2 Sentiment Analysis :

In Hindi, there are numerous open source datasets for sentiment analysis divided into three categories: "positive," "negative," and "neutral," however similar datasets are not accessible in other Indian languages. As a result, we've decided to take a large dataset in Hindi with data from multiple domains such as tweets, news articles, etc., and use the GoogleTrans module to translate a portion of it into other Indian languages, which we'll utilise for testing. We used the IIT Patna Hindi movie reviews, Hindi product reviews dataset and the IIIT Hyderabad Telugu movie reviews dataset for sentiment analysis. By applying a threshold rating of 3, the Tamil movie ratings dataset from Kaggle was turned into a sentiment analysis dataset.

3.2 NER :

In the majority of the experiments, we used FIRE 2013 NER datasets in Hindi, Malayalam, Tamil, and Bengali languages, which contain numerous entity tags such as location, person, count, year, date, entertainment, disease, artefact, period, organisation, disease, month, time, day, quantity, and a few open source datasets as well. FIRE 2013 NER dataset is an annotated corpus of news articles from various domains. We pre-processed the data and took only a few entities as per the requirements of the experiment.

Chapter 4

Text Classification

Text classification is the process of categorising a text into a set of words. Text categorization can use NLP to automatically analyse text and then assign a set of predetermined tags or categories depending on its context. There are different types Text classification tasks such as sentiment analysis, topic detection, language detection, etc. Despite having a large number of native speakers, the majority of Indian languages are considered resource poor. In most domains, there are not enough datasets available. As a result, text categorization becomes difficult for Indian languages.

Multilingual text classification approaches (Salil Aggarwal et al.,2021) are essential for classifying data in several languages. Training a multilingual model saves us from having to train distinct models for each language, and it also aids the system's development through parameter sharing. Many commonalities underpin the wide diversity of Indian languages. Most Indian languages have converged to a great extent as a result of contact over thousands of years. Many words in these languages share the same root word and meaning. They do, however, utilise various scripts evolved from the ancient Brahmi script but correspondences between comparable characters across scripts can be established.

Salil Aggarwal et al.,2021 shows that a single multilingual model trained by using linguistic relatedness beats the baselines by a wide margin and the model performs best when the vocabulary overlap between the language datasets is maximum. Andraž Pelicon et al., 2020 explored the performance of multilingual BERT based model on the sentiment analysis task on Slovenian news and evaluated its zero shot cross-lingual capabilities in Croatian language. Cindy Wang et al., 2021 presented an empirical evaluation of transformer-based text categorization models in a number of monolingual and multilingual pretraining and fine-tuning setting and showed that multilingual language models can outperform monolingual ones in some downstream tasks and target languages.

Some domain specific studies on multilingual data (Stephen Mutuvi et al., 2020) show that the models based on fine-tuned language models achieve very good performance on the classification of multilingual epidemiological texts for both high and low resource languages. Samujjwal Ghosh et al., 2022 proposed a graph neural network enhanced language models for disaster related multilingual text classification which works for multiple languages. Irene et al., 2021 proposed a general model agnostic framework for improving crosslingual text classification by leveraging the source instance weighting.

4.1 News Article Genre Classification

The task is to sort a news article into one of several categories, including sports, entertainment, business, politics, technology, and lifestyle. The aim of this experiment is to assess the performance of mBERT and XLMR on data from multiple Indian languages, both with and without fine-tuning, as well as to investigate the models' zero-shot cross-lingual

transfer learning capabilities under diverse conditions. For the majority of the tests, we used IndicNLP classification datasets as well as a few open source datasets from Kaggle..

Results of the experiments:

- Model used - mBERT base

1. On Telugu(3 classes - Entertainment, Business and Sports):
   (a) Without fine-tuning - 31.0%
   (b) After fine-tuning - 92.37%
2. On Tamil(3 classes - Entertainment, politics and Sports):
   (a) After fine-tuning - 94.35%
3. On Marathi(3 classes - Entertainment, Lifestyle and Sports):
   (a) After fine-tuning - 93.71%
4. Trained on Tamil and subsequently on Gujarati(3 classes - Business, Entertainment and Sports):
   (a) Tested on:
      i. Tamil - 93.75%
      ii. Gujarati - 91.62%
      iii. Telugu - 37.09%
5. Fine-tuned on Malayalam dataset which has 4 classes(business, entertainment, sports, technology) with different models:
   (a) BERT - 54%
   (b) mBERT - 90.05%
   (c) XLM-R - 90.83%

- Model used - XLMR base

1. Fine-tuned on Malayalam dataset(4 classes - Business, Entertainment, Sports and technology), test accuracies on:

(a) Telugu (business, sports and entertainment) - 90.996

(b) Gujarati (business, sports and entertainment) - 90.14

2. Fine-tuned on both Telugu and Tamil data

   - Telugu(5 classes) : “Business”, “editorial”, “entertainment”, “sports”, “nation”;
   - Tamil(6 classes) : ”tamilnadu”, ”india”, ”cinema”, ”sports”, ”politics”, ”world”.

   (a) Results :

      - Telugu- 95.44%
      - Tamil - 84.34%
      - Gujarati - 82.26
      - Malayalam - 88.38%
      - Bengali - 82.34%
      - Hindi - 91.62%
      - English - 87.32

Inferences:

- The performance of a pre-trained Multilingual language model is poor when tested without any fine-tuning.
- An mBERT model, after fine-tuning with data in a particular language, works pretty well for that specific language but cannot be extended to other languages.
- The XLM-R model fine-tuned monolingual dataset in one language works fine on other languages as well when the classes of test dataset are already seen during training.
- XLM-R specific to the dataset and domain is showcasing excellent cross-lingual transfer learning properties on Indian languages.

4.2 Sentiment Classification

The results of the News article genre classification task revealed that XLMR has remarkable cross-lingual skills. The aim of this experiment is to validate the findings of the previous experiment by using data from multiple domains to perform sentiment analysis. The task is to collect labelled sentiment analysis datasets from various domains like as tweets and product reviews in multiple Indian languages, train an XLM-R model using a monolingual dataset from a single domain, then test it across languages and domains. In Hindi, there are various open source datasets for sentiment analysis with three classes: "positive," "negative," and "neutral," however analogous datasets in other Indian languages are very few. As a result, we decided to take a huge dataset in Hindi and utilise the GoogleTrans module to translate a portion of it into various Indic languages for testing.

Results of the experiment:

- Model used - XLMR base

1. Fine-tuned on Hindi dataset with "positive", "negative" and "neutral" classes
2. Results :
    - Hindi - 91.93%
    - Malayalam - 77.54%
    - Tamil - 80.08%
    - Telugu - 79.61%
    - Bengali - 76.25%
    - Gujarati - 80.84%
    - Marathi - 83.91%

Inferences :

- Transfer learning properties were well exhibited in this experiment which validates that like in the News classification experiment which has intrinsic parallel property, this exp where we ensured parallel data through translation showcased positive result.
- The 10-15% drop in the test accuracy on languages other than Hindi can be attributed to the translation errors accumulated while preparing test datasets using googletrans library.
- This validates that XLM-R can be used to build multilingual capabilities in classification models if the data we are dealing with has parallel property.

4.3 Movie Reviews Classification

This experiment is part of a sentiment analysis study, although it is focused on movie reviews. We used the Hindi movie reviews dataset from IIT Patna, the Telugu movie reviews dataset from IIIT Hyderabad, and the Tamil movie ratings dataset from Kaggle and evaluated the zero-shot performance of XLMR model fine-tuned on a single language dataset.

Results of the experiment:

- Model used - XLMR base

1. Fine-tuned on Hindi dataset with 3 classes(positive, negative and neutral) and tested on Hindi only.
    - Test accuracy : 44.66
2. Fine-tuned on Telugu dataset with only 2 classes(positive and negative) and test accuracy are as follows:

    - Telugu - 80%
    - Hindi - 57.53%
    - Tamil - 59.50%

Inferences :

- The model fine-tuned on Hindi reviews is not performing well on Hindi data itself but the performance was good when fine-tuned with telugu language dataset from a different source.
- This experiment is not conclusive, we need to experiment more on this domain.

4.4 Extended Experiments

The goal of these extended experiments is to make minor tweaks to the previous experiments and investigate the cross-lingual capabilities of XLMR model with the effect of fine-tuning on multiple languages and variation of data.

Results of the experiment:

- Model used - XLMR base

1. Trained on Hindi and Telugu combined Sentiment Analysis dataset :
   - Data : Hindi - 4000 rows + Telugu - 222 rows (translated from hindi)
   - Test accuracy on:
     - Hindi(2269 rows) - 90.13%
     - Telugu(260 rows) - 83.85%
     - Gujarati(214 rows) - 80.84%
     - Marathi(219 rows) - 85.21%
     - Bengali(230 rows) - 79.70%
     - Tamil(226 rows) - 82.30%

   Inferences :

   - Test accuracy on Hindi got slightly decreased and around 2-4% of increase can be seen on other languages..

2. Product Reviews Sentiment Analysis :

- Trained on Hindi Product Reviews dataset(published by IIT Patna) with only two classes namely positive and negative.
  - Test accuracy:
    - Hindi : 92.88%
    - Telugu : 88.20%
- Trained on Telugu dataset (by IIIT Hyd) with positive and negative classes
  - Tested results:
    - Telugu product reviews - 83.87
    - Hindi Product reviews (by IIT Patna) - 91.90
    - Hindi sentiment analysis dataset (by IIT Patna) - 81.3
    - Telugu sentiment analysis dataset(translated from hindi dataset) - 73.71
    - Hindi movie reviews dataset(by IIT Patna) - 74.42
    - Gujarati Sentiment analysis (translated from Hindi dataset) - 77.483

Inferences :

- The result of the first experiment is as expected but in the second experiment, the XLM-R model fine-tuned on product reviews dataset in Telugu performed well on Hindi movie reviews dataset, which is not correlated to product reviews in any way.
- The evidence is inconclusive, we need to carry out more experiments on this aspect.

Chapter 5

Named Entity Recognition

The objective of the following experiments is to analyze the cross lingual transfer learning capabilities of XLMR model in Indic languages for NER task. In most of the experiments, we have used FIRE 2013 NER datasets. We have also used a few open source datasets from Kaggle.

Results of the experiment:

- Model used - XLMR base

1. Train on Telugu data with "name", "location", "organization", "misc" and "other" tags. The F1 score on test data is as follows:
   - On Telugu - 0.893
   - On Hindi(translated from telugu data using googletrans library) - 0.732

     Inferences :
     - This result shows a sign of transfer learning capability of XLMR on NER tasks.
2. Trained on FIRE 2013 Hindi data with "name", "location", "organization" and "other" tags, converted all other entity tags into "other" tag and refined the data by removing sentences that contain only "other" tags to reduce the skewness in the data used for fine-tuning. The F1 score on test data is as follows:

- On Hindi - 0.638
- On Tamil - 0.610
- On Malayalam - 0.590

Inferences :

- These results show that the XLMR model exhibits cross lingual capabilities but the scores are not as high as expected. Further experimentation by doing some pre-processing on train data is needed to validate these results.

3. Trained on FIRE 2013 Malayalam data with "person", "location", "organization", "cardinal", "time" and "others" tags. The "cardinal" and "time" tags are created by combining multiple tags provided in the original FIRE 2013 data. The results are as follows:
   - On Malayalam - 0.690
   - On Hindi - 0.602
   - On Tamil - 0.537

     Inferences :
     - The training data has a lesser number of sentences containing "organization" tag compared to other tags. Hence, most of the "organization" entities in test data were wrongly tagged as "others" which effected the overall performance of the model.
4. Trained on Malayalam data with "person", "location", "cardinal", "time" and "others" tags. The results are as follows:
   - On Malayalam - 0.695
   - On Hindi - 0.596
   - On Tamil - 0.543

- On Telugu - 0.625

  Tested only on sentences with number of "others" tag in the sentence lesser than 75% of the total number of tokens in the sentence.

- On Malayalam - 0.712
- On Hindi - 0.683
- On Tamil - 0.619
- On Telugu - 0.673

  The weighted F1 score computed by not considering the "others" tag is as follows:

- On Malayalam - 0.832
- On Hindi - 0.766
- On Tamil - 0.766
- On Telugu - 0.769

  Inferences :

  – If the data is highly skewed with comparatively huge number of "others" tag, the model performs poorly on test data.

  – The cross-lingual performance of the model significantly improved after removing sentences with more than 75% of "others" tag.

  – The weighted F1 scores computed by avoiding "others" tag are significantly better which shows that the overall F1 score is low due to huge number of tokens with "others" tag compared to other entities.

Chapter 6

Conclusions and Future work

In this paper, we explored the zero-shot cross-lingual transfer learning capabilities of pre-trained multilingual language models such as mBERT, XLMR in Indain languages. We conducted experiments in multiple NLP tasks such as News Article Genre Classification, Sentiment Analysis and Named Entity Recognition(NER). The text classification results show that XLM-R performs extremely well in a zero-shot setting when fined tuned on any Indian language and tested on any other language which is unseen during fine-tuning. The performance of the model is even better when there is an inherent parallelism between train data and test data. In NER task, XLMR showcases cross lingual capabilities but to a limited extent and the results improved under certain conditions of data used for fine-tuning. Further experimentation with rich annotated data is needed to draw rigid conclusions about the performance of XLMR in NER tasks.

Bibliography

[1] Salil Aggarwal, Sourav Kumar, and Radhika Mamidi. Efficient multilingual text classification for indian languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 19–25, 2021.

[2] Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*, 2016.

[3] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019.

[4] Gábor Berend. Massively multilingual sparse word representations. In *International Conference on Learning Representations*, 2019.

[5] 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.

[6] Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M Khapra, and Pratyush Kumar. Indicbart: A pre-trained model for natural language generation of indic languages. *arXiv preprint arXiv:2109.02903*, 2021.

[7] 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.

[8] Sumanth Doddapaneni, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra. A primer on pretrained multilingual language models. *arXiv preprint arXiv:2107.00676*, 2021.

[9] Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686*, 2018.

[10] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*, 2020.

[11] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, 2020.

[12] Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*, 2021.

[13] Yash Khemchandani, Sarvesh Mehtani, Vaidehi Patil, Abhijeet Awasthi, Partha Talukdar, and Sunita Sarawagi. Exploiting language relatedness for low web-resource language model adaptation: An indic languages study. *arXiv preprint arXiv:2106.03958*, 2021.

[14] Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2832–2838, 2017.

[15] Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit, and Huzefa Rangwala. Cross-lingual text classification of transliterated hindi and malayalam. *arXiv preprint arXiv:2108.13620*, 2021.

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

[17] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[18] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[19] Irene Li, Prithviraj Sen, Huaiyu Zhu, Yunyao Li, and Dragomir Radev. Improving cross-lingual text classification with zero-shot instance-weighting. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 1–7, 2021.

[20] Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. Choosing transfer languages for cross-lingual learning. *arXiv preprint arXiv:1905.12688*, 2019.

[21] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Mul-

tilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[22] 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.

[23] Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. What happens to bert embeddings during fine-tuning? *arXiv preprint arXiv:2004.14448*, 2020.

[24] Stephen Mutuvi, Emanuela Boroş, Antoine Doucet, Adam Jatowt, Gaël Lejeune, and Moses Odeo. Multilingual epidemiological text classification: a comparative study. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6172–6183, 2020.

[25] Andraž Pelicon, Marko Pranjić, Dragana Miljković, Blaž Škrlj, and Senja Pollak. Zero-shot learning for cross-lingual news sentiment classification. *Applied Sciences*, 10(17):5993, 2020.

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

[27] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[28] Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification. *arXiv preprint arXiv:1908.11860*, 2019.

[29] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[30] Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. Cross-lingual transfer learning for multilingual task oriented dialog. *arXiv preprint arXiv:1810.13327*, 2018.

[31] Muhammad Haroon Shakeel, Asim Karim, and Imdadullah Khan. A multi-cascaded deep model for bilingual sms classification. In *International Conference on Neural Information Processing*, pages 287–298. Springer, 2019.

[32] Lei Shi, Rada Mihalcea, and Mingjun Tian. Cross language text classification by model translation and semi-supervised learning. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1057–1067, 2010.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[34] Zihan Wang, Stephen Mayhew, Dan Roth, et al. Extending multilingual bert to low-resource languages. *arXiv preprint arXiv:2004.13640*, 2020.

[35] Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*, 2019.

[36] Shijie Wu and Mark Dredze. Are all languages created equal in multilingual bert? *arXiv preprint arXiv:2005.09093*, 2020.

[37] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A

massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

[38] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*, 2019.

[39] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[40] Ziqing Yang, Yiming Cui, Zhigang Chen, and Shijin Wang. Cross-lingual text classification with multilingual distillation and zero-shot-aware training. *arXiv preprint arXiv:2202.13654*, 2022.