Explainability in Low-Resource and Multilingual NLP Applications

KULDEEP LODHA

Research Proposal

NOVEMBER 2024

ABSTRACT

Multilingual natural language processing (NLP) advancements have simplified overcoming language barriers through models like BERT and XLM-R. However, these models are still difficult to understand in low-resource languages, which makes them less trustworthy and harder to use in important areas like healthcare, legal analysis, and education.

Our research focuses on making multilingual models more understandable in low-resource settings. We use explainable AI (XAI) techniques, especially SHAP (Shapley Additive Explanations), LIME, and attention visualization, to improve the transparency of the model's predictions by using multilingual datasets. We aim to show that explainability can improve trust and model performance in low-resource settings. Additionally, we aim to find a balance between computational efficiency and understandability, which is important for low-resource settings with limited hardware.

This research aims to close the gap between advanced multilingual NLP models and real-world applications in low-resource languages. The results of the study are expected to provide a better understanding of model prediction to the user and provide a roadmap for future research in explainable multilingual NLP.

LIST OF ABBREVIATIONS

NLP	Natural Language Processing
SHAP	Shapley Additive Explanations
LIME	Local Interpretable Model-Agnostic Explanations
CPU	Central Processing Unit
GPU	Graphical Processing Unit
TPU	Tensor Processing Unit
MBERT	Multilingual Bidirectional Encoder Representations from Transformers
XLM-R	Cross-lingual Language Model
RAM	Random Access Memory
ROUGE	Recall-Oriented Understudy for Gisting Evaluation

LIST OF FIGURES

Table of Contents

1. Background

Over the past few decades, Natural Language Processing has witnessed outstanding progress by advances in machine learning and the availability of a wide range of text corpora. Some multilingual models such as mBERT and XLM-R have emerged as important tools, enabling language understanding across diverse linguistic contexts. These models are trained to support many languages in a single framework for tasks like translation, sentiment analysis, and named entity recognition. These models work satisfactorily for high-resource languages, where a variety of training data is available. On the other hand, for low-resource languages, these models underperformed which leads to a significant gap in NLP's inclusivity.

Now let's understand the strategy behind the working of these models like how they decide which words are most important. Why do they fail in specific linguistic contexts? This is where the field of explainability has come into the picture. Explainability in AI systems has become an important part as machine learning models become more complex. This addresses all the questions with the help of techniques such as SHAP, LIME, and attention visualization and provides insights into how models make predictions. With the help of these methods, we can understand the process and strategy behind the working of models, debug models, and promote the ethical use of AI. However, these techniques are primarily tested on high-resource language tasks and rarely tested on multilingual and low-resource settings. There are challenges like cultural differences, language biases, and tokenization problems, which make models harder to understand.

This study focuses on exploring and addressing these challenges with the help of explainability techniques for multilingual NLP models in low-resource settings. We aim to make NLP systems easy to understand, fair, and accessible for everyone by linking model transparency with language inclusivity.

2. Problem Statement

Multilingual models like mBERT and XLM-R are great tools for language processing, but they still face significant challenges for low-resource languages due to limited resource availability. While these models work well to capture linguistic nuances in the high-resource settings their accuracy drops significantly for low-resource settings. This problem becomes bigger due to a lack of unclarity, how do these models make predictions?

Explainability techniques (e.g. SHAP and LIME) answer all these questions and provide deep insights into the model's predictions. These techniques are required for debugging and trust-building in AI development and are designed for high-resource languages. Explainability techniques such as LIME and SHAP have been widely used to interpret deep learning NLP models. They have effectively identified important words and detected biases in tasks like sentiment analysis and text classification(Lundberg & Lee, 2017; Ribeiro et al., 2016). However, these methods often face difficulty in multilingual settings because tokenisation and meaning vary between languages.

Deep learning Models like mBERT(Devlin et al., 2018) and XLM-R(Conneau et al., 2019) are a good step forward in multilingual learning. They use shared representation to work on multiple languages simultaneously. However, (Wu & Dredze, 2020)research indicates that these models struggle to perform well in low-resource languages where their accuracy drops significantly.

There are some recent research shows that explainability is important for making AI systems fair and transparent. However:

1. Most explainability techniques are tested in English and other high-resource languages (Danilevsky et al., 2020).

2. explainability techniques remain unexplored in Multilingual NLP models for low-resource settings (Hangya et al., n.d.).

3. The impact of tokenization on interpretability in multilingual tasks is less explored (Rust et al., 2021).

3. Research questions

The research questions suggested for each research for every research objective are as follows:

1. How may explainability techniques help improve the interpretability of complicated multilingual NLP models in low-resource settings?

2. What are the primary issues associated with using explainability methods in low-resource languages, and how do we handle them?

3. What is the difference in accuracy and computing cost across different explainability methods in low-resource settings?

4. What are the main trade-offs between interpretability, accuracy, and computational efficiency when using multilingual NLP models in low-resource settings?

4. Aim and Objectives

This research focuses on examining the background of the Multilingual NLP models for low-resource languages and investigating the existing methods of Explainable AI that can be used in low-resource environments. There are different methods that can be used but there is a lack of study concerning the feasibility and effectiveness of those models to make more relevant and trustworthy predictions.

The objectives of the research are as follows: -

1. To explore existing multilingual models like mBERT and XLM-R and the different challenges associated with the multilingual models when we use them in low-resource settings.

2. To explore the different explainable AI methods and develop a systematic pipeline to apply them in multilingual tasks.

3. To perform experiments on a multilingual dataset and evaluate the impact of explainable AI on model performance.

4. To improve the efficacy of explainability techniques in low-resource settings. Investigate easier methods for improving accuracy, transparency, and processing power. Also, explore ways to reduce processing requirements.

5. To identify limitations and challenges of using explainability methods with multilingual models.

6. Propose the guidelines for integrating explainable AI methods in low-resource settings.

The goal of this research is to use explainable AI techniques that enhance model transparency, accuracy, and usability while ensuring computational efficiency to improve the clarity of multilingual NLP models, particularly for low-resource languages.

5. Significance of the study

This research is important in multilingual natural language processing as it addresses key points in understanding how multilingual models like mBERT and XLM-R work, particularly in low-resource settings. It aims to make a more efficient and effective way to make model predictions easy to understand and trustworthy with the help of explainable AI techniques. This study's outcome will enhance trust and efficiency, facilitating the fair and responsible use of NLP systems across various languages.

6. Scope of the study

This study will use Explainable AI to better understand how multilingual NLP models like mBERT and XLM-R operate, especially in low-resource settings. This includes reviewing and analyzing the accuracy of various explainability strategies and techniques (such as LIME, SHAP, and attention-based approaches) in predicting the model's predictions. The study will be focused on real-life scenarios and practical research.

7. Research Methodology

The research methodology involves advanced XAI techniques and how they can be used to handle the difficulties of multilingual and low-resource environments. This includes selecting a model, fine-tuning it, and applying explainability methods to clarify the logic behind its predictions. Evaluation measures will be developed to evaluate interpretability, model performance, and user trust.

The steps involved in research methodology are as follows:

7.1. Review of the literature: -

Start a comprehensive review of already available literature on multilingual NLP models (e.g., mBERT, XLM-R) and explainability methods (e.g., SHAP, LIME) and related topics.

7.2. Dataset Selection and Preprocessing: -

Dataset selection: - Identity multilingual datasets that include both high and low-resource languages. For this study, the datasets to be used are: -wikiann, flores_101, mafand

Data Preprocessing: the steps of data preprocessing are as follows: -

- Tokenization: Apply the Tokenization on the dataset and ensure low-resource languages are tokenized without losing semantic context.
- Special Text Removal: Remove special characters or symbols that the language model hard to recognize. For example, comments, string literals, and special characters. They need some extra processing to handled properly.
- Removing Noise: Remove noisy or irrelevant elements from the text to ensure high-quality inputs for training and analysis.
- Normalization and Standardization: Normalization and standardization are the essential parts of data preprocessing, especially in multilingual NLP tasks. It ensures that the text data is consistent, reducing variability that could confuse the models.
- Data Cleaning and Augmentation: Clean the data to improve consistency by removing duplicates, and incomplete sentences. Additionally, use augmentation to balance the dataset with additional training data to increase performance of model.
- Length Limitation: Handle input length limitation by cutting long text into small parts that fit as per the input size.

7.3. Model Selection:

In this section, we are going to select appropriate models and establish baseline methods to evaluate their performance.

7.3.1. State-of-the-Art Models:

- XLM-R: We have chosen this model for its ability to handle cross-lingual tasks effectively, particularly in low-resource settings, as it is pre-trained on a wide range of multilingual corpus.
- mBERT: - We have selected this model for simplicity and accessibility, serving as a baseline model for comparison with more recent models.

7.3.2. Explainability Add-Ons:

To make the model easier to understand, Explainability techniques such as SHAP and LIME will be used.

7.3.3. Baseline Models:

- To establish initial benchmarks, Traditional models, including TF-IDF + SVM will be used.
- To evaluate multilingual performance cross-lingual embeddings like LASER will serve as an alternative baseline.

7.4. Training: -

The training process is divided into two parts:

- Fine-Tuning: We are going to fine-tune the pre-trained multilingual models (such as mBERT, XLM-R) for task-specific datasets. This involves applying transfer learning to minimize computational requirements while maintaining accuracy.
- Explainability Model Integration: After training, we will integrate Explainability techniques such as SHAP and LIME to generate insights from model predictions.

7.5. Evaluation:

Evaluating the model's predictions is an essential part of determining the validity and effectiveness of the proposed approach. The metrics that will be utilized can vary depending on the distinct task and application. Below are the metrics that will be utilized.

- Accuracy: Accuracy Measures the ratio of correct predictions out of the total predictions. These metrics are used in classification tasks like sentiment analysis to evaluate the language model's ability to generate accurate predictions.
- Perplexity: Perplexity metrics evaluate the model's ability to predict a sequence of text. A lower perplexity score means superior performance. This is widely used metric to assess the model's text generation capabilities.
- ROUGE Score: The ROUGH score will be used to measure the quality of text summarization. It evaluates how much the generated summary matches with the reference summary by considering recall, precision and F1 score.
- Human Evaluation: It involves the evaluation of the generated text or output by human judgement. This includes checking how fluent the text is, how relevant it is, and grammatical accuracy and then giving an overall quality rating.

7.6 Analysis:

Analysis of the results to conclude and identify insights regarding the feasibility and effectiveness of using explainability techniques on advanced NLP models in multilingual NLP, especially in low-resource languages.

8. Requirements Resources

This research will require access to the necessary resources, tools and datasets. It includes hardware, software, dataset and any specific requirements for the study.

8.1 Hardware Requirements

Advanced NLP models like mBERT and XLM-R require a high-performance computing environment for training and fine-tuning. These resources typically include:

- High-performance: To speed up the Training process of deep learning models requires with high-performance GPU resources.
- Special hardware accelerators: Some platforms like (Kaggle) use TPUs or other environments to boost the training.
- RAM: Model training requires memory resources for storing datasets, models, and intermediate outputs.

8.2 Software Requirements

The software requirements for training Advance NLP models include:

- Deep learning frameworks: TensorFlow, PyTorch, NLP, or similar packages are commonly used for tokenization, data pre-processing and training models.
- Efficient data processing libraries: Tools like Pandas and Scikit-learn can help to manage large datasets efficiently.
- Hugging Face Transformers: It is an open-source library which contains pre-trained models and tools, widely used for NLP tasks.

9. Research Plan

Week 1-2: Research proposal development

- Week 1: Identify the research objective, scope of the study, and Methodology.
- Week 2: Start working on the literature review for the research proposal.

Week 3 - 4: Research proposal review

- Week 3: Send the proposal for review to the thesis supervisor.
- Week 4: Based on the feedback provided revise the proposal and make submission.

Week 5 - 8: Data Collection and Interpretation

- Week 5-7: Collect data using surveys, experiments or other Methods
- week 8: Start the data analysis using different statistical methods

Week 9-10: Research Initial Draft and Review

- Week 9: Start working on the initial draft of the research, which includes the sections Abstract, Introduction, Research questions, Dataset Description, and Methodology.
- Week 10: Formatting report and submitting it for initial review by the supervisor.

Week 11-13: Data Preprocessing and Initial Model Development

- Week 11: Preprocessing of data.
- Week 12: Start working to develop the initial model and training.
- Week 13: Model Evaluation and applying model optimization techniques.

Week 14-15: Model Refinement and Evaluation

- Week 14: Fine-tuning the model and integrating explainability techniques.
- Week 15: Validate initial results with test datasets and compare the performance of proposed methods for evaluation.

Week 16-17: Finalize Model and Results:

- Week 16: Finalize the best model based on the performance.
- Week 17: Draft Model result for research report

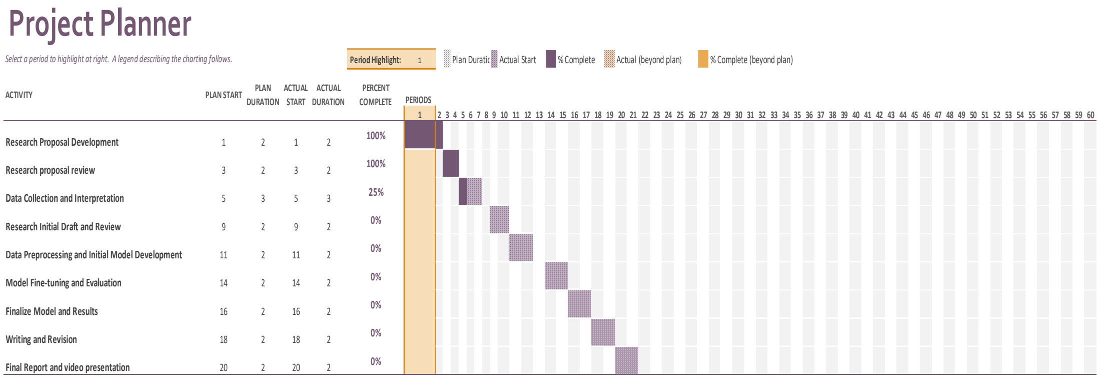
Week 18-19: Writing and Revision

- Week 18: Prepare the final report and send it for the review.
- Week 19: Based on the feedback, Revise the final report.

Week 20-21: Presentation Preparation

- Week 20: Start working on the video presentation.
- Week 21: Finalize Research and Submission
- Week 21: Finalize the final report with a video presentation and submit it.

Here is the Gantt chart for the research timeline:



Project Planner

Select a period to highlight at right. A legend describing the charting follows.

Period Highlight: 1

▨ Plan Duratic ▨ Actual Start ■ % Complete ▨ Actual (beyond plan) ■ % Complete (beyond plan)

ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Research Proposal Development	1	2	1	2	100%
Research proposal review	3	2	3	2	100%
Data Collection and Interpretation	5	3	5	3	25%
Research Initial Draft and Review	9	2	9	2	0%
Data Preprocessing and Initial Model Development	11	2	11	2	0%
Model Fine-tuning and Evaluation	14	2	14	2	0%
Finalize Model and Results	16	2	16	2	0%
Writing and Revision	18	2	18	2	0%
Final Report and video presentation	20	2	20	2	0%

Figure 1 - Project Plan

References

1. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. and Stoyanov, V., (2019) Unsupervised Cross-lingual Representation Learning at Scale.

2. Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y. and Sen, P., (2020) *A Survey of the State of Explainable AI for Natural Language Processing*. [online] Available at: https://xainlp2020.github.io/xainlp/.

3. Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K., (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

4. Hangya, V., Saadi, H.S. and Fraser, A., (n.d.) *Improving Low-Resource Languages in Pre-Trained Multilingual Language Models*. [online] Available at: https://cistern.cis.lmu.de/lowresCCWR.

5. Lundberg, S. and Lee, S.-I., (2017) A Unified Approach to Interpreting Model Predictions.

6. Ribeiro, M.T., Singh, S. and Guestrin, C., (2016) 'Why Should I Trust You?': Explaining the Predictions of Any Classifier.

7. Rust, P., Pfeiffer, J., Vuli´cvuli´c, I., Ruder, S. and Gurevych, I., (2021) *How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models*. [online] Available at: www.ukp.tu-darmstadt.de.

8. Wu, S. and Dredze, M., (2020) *Are All Languages Created Equal in Multilingual BERT?*