Multilingual Named Entity Recognition (NER) for Switzerland

Ahmed Elzayat, Mohamed Alaa, Youssef Ahmed, Zeyad Ahmed

Abstract*—This paper explores the use of multilingual transformers, specifically mBERT and XLM-RoBERTa, for named entity recognition (NER) in the context of Switzerland's multilingual environment. Utilizing the PAN-X dataset, which encompasses German, French, Italian, and English, we examine both multilingual and monolingual fine-tuning strategies. Our findings reveal that mBERT consistently outperforms XLM-RoBERTa in both cross-lingual and monolingual settings, offering a robust solution for Switzerland's linguistic diversity. This challenges existing literature that highlights XLM-RoBERTa's advantages due to its extensive pretraining corpus. We discuss potential reasons for this discrepancy, including dataset characteristics and model architecture. This study underscores mBERT's effectiveness for NER tasks in multilingual contexts and provides valuable insights for optimizing transformer models to address the unique linguistic challenges in Switzerland. *Source code: https://github.com/MohamedAlaaAli/SWIZT/tree/main

***Index Terms*—Multilingual Transformers, Named Entity Recognition (NER), mBERT, XLM-RoBERTa, Cross-lingual Transfer.**

I. Introduction

NAMED Entity Recognition (NER) is widely used and a critical use of Natural language Processing (NLP). It identifies and categorizes entities as persons, locations as well as organizations within a piece of text. There has been a lot of research made developing a variety of NER systems that rely on a single language.

The primary goal of NER systems is to reduce the time and effort spent on tasks such as document analysis, search optimization, and automated summarization. For example, in businesses, NER can quickly extract entities from contracts or news articles that are relevant to the business, which speeds up decision-making processes as it summarizes keynote entities involved. Similarly, in healthcare, it can help identify critical patient information from medical records automatically, enhancing efficiency and accuracy. This makes NER a vital application of NLP due to the ability to automate entity identification and its use in many applications.

However, the task becomes dramatically complex as we add more languages to the system making it multilingual. Multilingual transformers have shown great results becoming a promising solution to fill this gap, enabling cross-lingual transfer for many NLP problems. A critical point of these models is the ability to generalize languages without the need to differentiate or distinguish between explicit languages. Eliminating the need to train for specific language models.

One of the known Multi-lingual transformers is Multilingual BERT (mBERT), this model refines the architecture and objectives of the BERT transformer by incorporating multi-lingual corpora during pretraining. mBERT reveals superb performance in transferring knowledge across several languages, making it effective in many multi-lingual NLP tasks, including NER. However, mBERT's performance could be limited in languages that are not spoken often due to constraints in its training corpus size.

Another multi-lingual transformer is XLM-RoBERTa (XLM-R), this model has shown significant advances due to a much larger pertaining corpus, including Wikipedia dumps and 2.5 terabytes of Common Crawl data, spanning 100 languages. This global dataset allows the XLM-R to slightly outperform mBERT, particularly in inferior languages. Moreover, XLM-R adopts SentencePiece tokenization, preserving subword structures and ensuring effective handling of multi-lingual text without relying on any specific language pre-processing to the input sequences.

In this study, we explore the application of different multilingual transformers, particularly XLM-R and mBERT to NER tasks in the multi-lingual Swiss context, focusing on German, French, Italian and English languages. This will be done by employing the WikiANN (PAN-X) dataset [1]. which also uses the IOB2 tagging format for location, person and organization entities. we aim to simulate realistic test cases reflecting Switzerland's linguistic diversity. We also shed light on challenges like subword tokenization in NER by adopting conventional solutions established by BERT, allowing suitable compatibility with the IOB2 format.

II. Related Work

Named Entity Recognition (NER) in multilingual settings has become a prominent area of research due to the development of large-scale multilingual benchmarks, advances in transformer architectures, and innovative pretraining methods. These advancements were driven by several challenges related to low-resource languages, code-switching, and computational demand.

One of the early attempts to address the multilingual NER problem arose from the challenges posed by code-switching, where multiple languages appear within a single text. Solorio et al. (2014) [2] addressed this issue by focusing on NER in code-switched text and proposed models that integrated language identification modules into the NER pipeline. Their approach involved conditional random fields (CRFs) and support vector machines (SVMs) for sequence labeling, leveraging linguistic features such as part-of-speech tags and word-level language identification to handle the multilingual context.

The study demonstrated that incorporating language identification significantly improved the model's ability to process mixed-language text. However, their models struggled with highly complex code-switching scenarios, particularly when dealing with structurally divergent languages, such as those with different syntax or word-order conventions. This highlighted the need for more sophisticated methods capable of better handling the linguistic variability inherent in code-switched text.

A key resource for multilingual NER is the WikiANN benchmark, which offers annotated NER data for more than 40 languages and has become a standard for training and evaluating multilingual NER models. Studies by Hu et al. (2020) and Pan et al. (2017) [[1], [3]] highlighted that leveraging a large multilingual training dataset significantly enhances performance, especially for low-resource languages. Nonetheless, the WikiANN dataset is derived from Wikipedia text, which might not adequately capture the domain-specific terminology required for certain applications.

One of the most notable studies in multilingual NER is the work of Pires et al. (2019) [4] on multilingual BERT (mBERT) for NER tasks. mBERT, among the first multilingual transformers, shares the same architecture and pretraining objective as BERT but incorporates Wikipedia articles from multiple languages into its pretraining corpus.

In a study by Bhosale et al. (2020) [5], the authors specifically addressed the challenge of performing NER for low-resource languages. Their approach combined rule-based methods with machine learning techniques to improve NER performance for languages with limited annotated data. They proposed a hybrid method that used a pre-trained multilingual model, such as mBERT, in combination with data augmentation techniques to improve entity recognition in low-resource settings. While their approach showed promise, it also highlighted the difficulty in achieving high accuracy in such settings.

Following mBERT, XLM-RoBERTa (XLM-R) emerged as its successor. XLM-R relies solely on the MLM pretraining objective across 100 languages, but it stands out due to the massive size of its pretraining corpus. This corpus includes Wikipedia dumps for each language alongside 2.5 terabytes of Common Crawl web data, making it several orders of magnitude larger than those used in earlier models. This extensive dataset provides a significant advantage for low-resource languages such as Burmese and Swahili, which have a limited number of Wikipedia articles available (Conneau et al., 2020) [6].

The "RoBERTa" part of the model's name indicates that its pretraining approach mirrors that of monolingual RoBERTa models (Liu et al., 2019) [7]. RoBERTa introduced several improvements over BERT (Devlin et al., 2019) [8], most notably by eliminating the next sentence prediction task entirely. Similarly, XLM-R omits the language embeddings used in XLM and employs SentencePiece to tokenize raw text directly. In addition to its multilingual design, a key distinction between XLM-R and RoBERTa lies in the size of their vocabularies: XLM-R has 250,000 tokens, compared to RoBERTa's 55,000.

III. Dataset

This project utilizes the WikiANN (PAN-X) dataset [1], a multilingual corpus for named entity recognition (NER). The dataset is derived from Wikipedia articles and is annotated with three primary entity types:

- **LOC:** Locations (e.g., countries, cities, landmarks).
- **PER:** Persons (e.g., individual names).
- **ORG:** Organizations (e.g., companies, institutions).

The dataset supports 176 languages, making it a rich resource for evaluating multilingual and cross-lingual NER problems in NLP. Each language is divided into three subsets or splits, including train, validation and test, each consisting of three different fields:

- ***tokens:*** List of strings representing the tokens in the sentence.
- ***langs:*** List of string features indicating the language of each token.
- ***ner_tags:*** List of indices representing entity types for each token.

The list of NER indices can be mapped to their corresponding tags in the **IOB2** format, providing a more interpretable representation, as shown in Table I.

Index	Entity Type
0	`O` (Outside any entity)
1	`B-PER` (Beginning of a person entity)
2	`I-PER` (Inside a person entity)
3	`B-ORG` (Beginning of an organization entity)
4	`I-ORG` (Inside an organization entity)
5	`B-LOC` (Beginning of a location entity)
6	`I-LOC` (Inside a location entity)

TABLE I
MAPPING OF INDICES TO ENTITY TYPES (**IOB2 FORMAT**)

In addressing our Swiss problem, we will utilize the English (en), Italian (it), German (de), and French (fr) corpora to fine-tune our target models. The entire dataset is balanced across splits, ensuring an equal distribution of all entities within each split as shown in figure 1

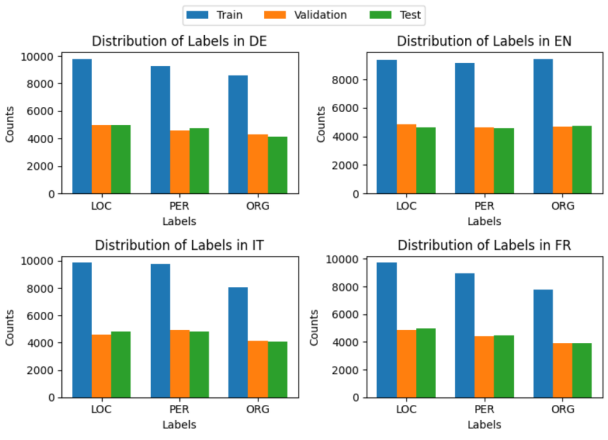

Fig. 1. Entity Distribution Across Splits in Each Language Corpus

IV. Methods

This study employs two state-of-the-art multilingual transformer models: **XLM-RoBERTa (XLM-R)** and **m-BERT**. These models are pre-trained on massive multilingual corpora, enabling effective cross-lingual transfer for NER tasks. The choice of these models is motivated by their ability to generalize well across languages without requiring separate models for each language. XLM-RoBERTa is a multilingual extension of RoBERTa, pre-trained using masked language modeling (MLM) on 100 languages. It employs SentencePiece tokenization and is trained on a larger corpus compared to earlier models, boosting its performance on low-resource languages. m-BERT extends BERT's architecture to multiple languages using MLM, with Wikipedia articles in multiple languages as its training data.

The network architecture for NER is framed as a token classification task:

1) **Input Layer:** Tokenized text using the SentencePiece tokenizer for XLM-R and WordPiece tokenizer for m-BERT.
2) **Encoder:** A transformer-based architecture (e.g. XLM-R or m-BERT), which outputs contextual embeddings for each token.
3) **Classification Head:** A linear layer maps token embeddings to label probabilities.
4) **Loss Function:** Cross-entropy loss, with a mask value (-100) for ignored tokens.

We additionally employed regularization of the hidden states using a dropout layer to prevent overfitting during fine-tuning. Fig. 2 illustrates the flow of our token classification architecture:

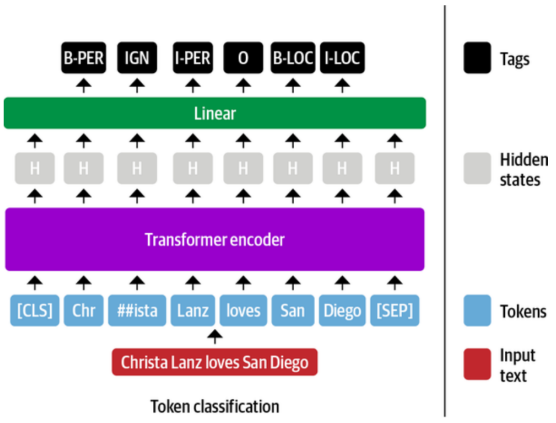

Fig. 2. NER Architecture with Transformer Models. Input text is tokenized and processed by a transformer encoder to generate hidden states (H). A linear layer maps these states to NER tags (e.g., B-PER, I-LOC, O) in the IOB2 format. Special tokens like [CLS] and [SEP] are ignored.

The loss function used for fine-tuning the XLM-RoBERTa and m-BERT models is the categorical cross-entropy loss, which is well-suited for multi-class classification problems like NER. In this task, the model assigns a probability distribution over possible labels (e.g., B-PER, I-PER, O) for each token in the input sequence. The cross-entropy loss quantifies the discrepancy between the predicted probability distribution and the true label distribution.

For a single token, the cross-entropy loss is defined as:

$$L_{\text{token}} = -\sum_{k=1}^{C} y_k \log(\hat{y}_k) \tag{1}$$

where C is the number of possible entity tags (e.g., B-PER, I-PER, O), y_k is the true label represented in one-hot encoding (1 for the correct label, 0 otherwise), $\hat{y}_k$ is the predicted probability for the k-th label, obtained using the softmax function:

$$\hat{y}_k = \frac{\exp(z_k)}{\sum_{j=1}^{C} \exp(z_j)} \tag{2}$$

Here, z_k represents the unnormalized logit output of the classification head for the k-th label. The loss for an entire input sequence is computed as the average loss across all valid tokens in the sequence:

$$L_{\text{sequence}} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{C} y_{t,k} \log(\hat{y}_{t,k}) \tag{3}$$

where T is the length of the tokenized sequence (excluding padding and ignored tokens), $y_{t,k}$ is the true label for the t-th token and k-th label, $\hat{y}_{t,k}$ is the predicted probability for the t-th token and k-th label.
To handle special tokens (e.g., [CLS], [SEP]), subword tokens, and padding tokens, a masking mechanism is applied. Tokens labeled as -100 are ignored during the loss computation using the mask m_t, where:

$$m_t = \begin{cases} 1 & \text{for valid tokens,} \\ 0 & \text{for ignored tokens.} \end{cases}$$

The final loss for NER is computed as:

$$L_{\text{NER}} = -\frac{1}{\sum_{t=1}^{T} m_t}\sum_{t=1}^{T} m_t \sum_{k=1}^{C} y_{t,k} \log(\hat{y}_{t,k}) \tag{4}$$

This approach ensures that only valid tokens contribute to the loss calculation, allowing the model to focus on predicting entity labels while ignoring irrelevant subword and padding tokens.

V. Results

For our task, both mBERT and XLM-RoBerta models were fine-tuned on the PAN-X dataset, which includes the four most widely spoken languages in Switzerland: *German, French, Italian, and English*. Fine-tuning was conducted in two settings:

- **Multilingual Fine-tuning:** Models were fine-tuned on all four languages simultaneously and tested on each language. This approach leverages shared linguistic features across languages to enhance performance.

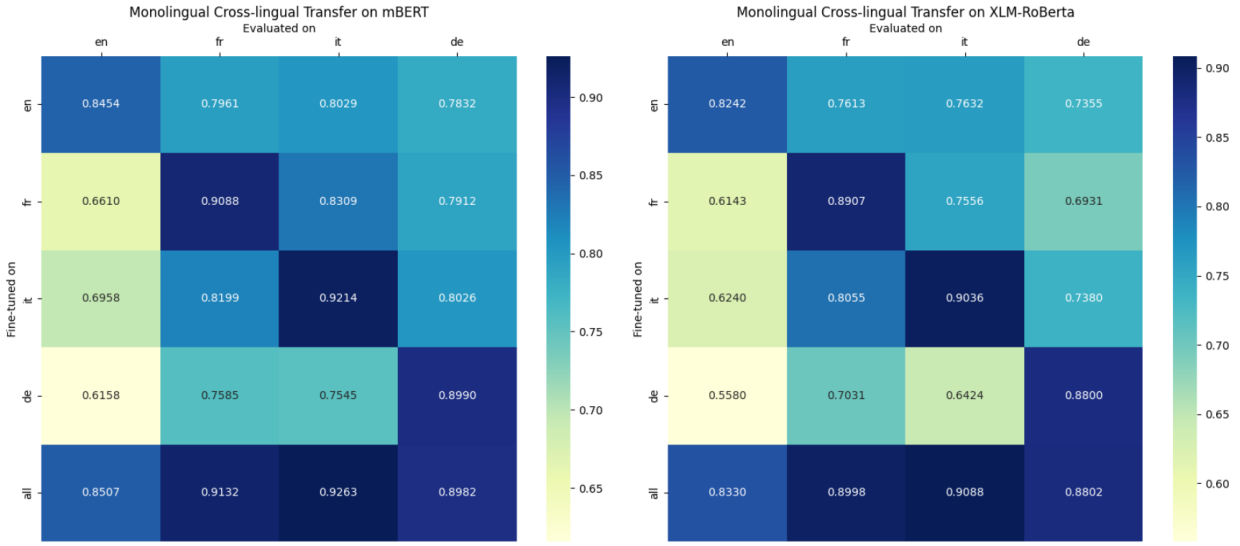

Fig. 3. Comparison of mBERT and XLM-RoBerta models on multilingual NER tasks. Each confusion matrix displays $F1$ scores for fine-tuning and evaluating on English (en), French (fr), Italian (it), and German (de). The 'all' row represents performance across all languages after fine-tuning on the whole languages collectively. Darker shades indicate higher $F1$ scores.

- **Monolingual Fine-tuning:** Models were fine-tuned on a single language and tested on the others to evaluate cross-lingual generalization capabilities. This setting assesses the model's ability to transfer learned representations to different linguistic contexts.

The results are visualized in the confusion matrices (Figure 3), which display the $F1$ scores for each language pair. The diagonal values represent the model's performance when fine-tuned and evaluated on the same language, while off-diagonal values indicate cross-lingual transfer performance. The 'all' row in each matrix summarizes the model's overall performance across all languages after being fine-tuned on whole set of languages.

The comparison between mBERT and XLM-RoBerta reveals that mBERT consistently outperforms XLM-RoBerta in both cross-lingual and same-language settings. mBERT achieves higher F1 scores across all language pairs and demonstrates superior generalization capabilities when fine-tuned on the entire dataset and evaluated on each language individually.

VI. Discussion and Summary

Our results indicate that mBERT outperforms XLM-RoBerta in both cross-lingual and monolingual NER tasks, which contrasts with some findings in the literature. Notably, Pires et al. (2019) [4] demonstrated the effectiveness of mBERT for multilingual NER, highlighting its ability to leverage multilingual data.

XLM-RoBerta, as described by Conneau et al. (2020) [6], was expected to excel due to its extensive pretraining corpus, which includes a vast amount of multilingual data. The model's design, which omits language embeddings and uses SentencePiece tokenization, was intended to enhance its performance across diverse languages.

The discrepancy between our findings and the literature could be attributed to several factors:

- **Dataset Characteristics:** The PAN-X dataset used in our experiments may have specific features that favor mBERT's architecture or pretraining approach.
- **Evaluation Metrics:** Differences in evaluation metrics or experimental setups could lead to varying results.
- **Model Fine-tuning:** The fine-tuning process, including hyperparameter settings and data preprocessing, might have influenced the models' performance differently.
- **Language Coverage:** While XLM-RoBerta's extensive corpus is advantageous, mBERT's architecture might be better suited for the specific languages and tasks in our study.

Further investigation is needed to understand these discrepancies fully. Future work could explore different datasets, fine-tuning strategies, and model configurations to reconcile these differences.

APPENDIX A
ANALYSIS OF PERFORMANCE

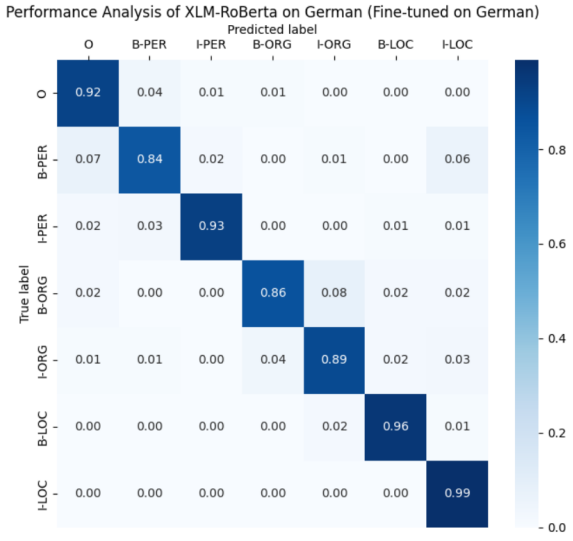

Fig. 4. Normalized confusion matrix illustrating the performance of XLM-RoBerta on the German NER task, with the model fine-tuned specifically on German data. As shown, the models gets confused between the *(B-ORG)* and *(I-ORG)*.

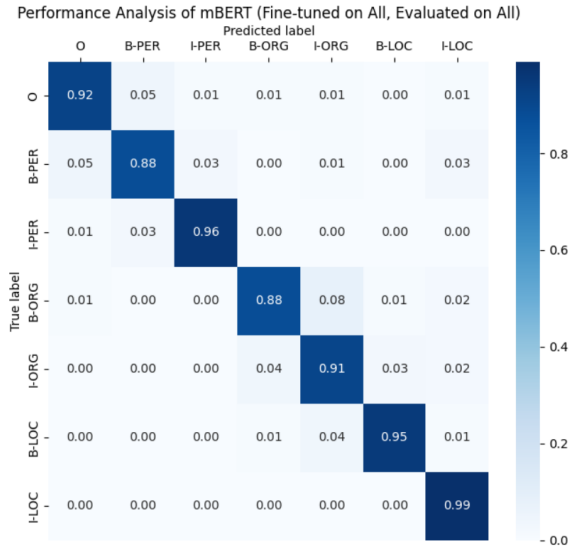

Fig. 5. Normalized confusion matrix illustrating the performance of mBERT on the multilingual NER task, with the model fine-tuned on all language corpora data. As shown, the models gets confused between the *(B-ORG)* and *(I-ORG)*.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Lewis Tunstall, Leandro von Werra, and Thomas Wolf for their book "Natural Language Processing with Transformers: Building Language Applications with Hugging Face," with a foreword by Aurélien Géron, which served as an essential resource throughout our work. We also thank Dr. Inas A. Yassine and our teaching assistant, Samar Alaa, for their invaluable support and guidance.

REFERENCES

[1] J. Hu *et al.*, "Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2020.

[2] T. Solorio *et al.*, "Overview of the third workshop on language technology for code-switching," in *Proceedings of the 3rd Workshop on Language Technology for Code-Switching*, 2014.

[3] X. Pan *et al.*, "Cross-lingual named entity recognition using wikipedia," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

[4] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?" *arXiv preprint arXiv:1906.01326*, 2019.

[5] S. Bhosale and A. Joshi, "Low-resource named entity recognition: A study on marathi and hindi," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

[6] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[7] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[8] J. Devlin *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.