University of Khartoum
Department of Electrical and Electronics Engineering

ENIGMA

SudaBERT: A Pre-trained Encoder Representation For Sudanese Arabic Dialect

Khalid N. Elmadani*, Mukhtar Elgezouli*, Muhammed Saeed*
*All authors contributed equally

Objectives

- Build a Natural Language Understanding (NLU) system for Sudanese Arabic dialect.
- Use the system to answer user's questions by matching them to a given set of previously answered questions from the database.

Introduction

- BERT is transformer-based NLU model. It takes as input a sentence and produces a contextual representation of it.
- There is an Arabic version of BERT called Arabic-BERT, Arabic-BERT is trained on OSCAR Arabic dataset.
- Although OSCAR contains huge amounts of Arabic text, the representation of Sudanese dialect there is poor.

Methodology

Data Collection

We've collected about 13 million Sudanese sentences from twitter and public Telegram channels. Then, we cleaned the data from all symbols, emojis and non-Arabic words. Finally, we ended up with about 7 million cleaned sentences.

Pre-training

- We tokenized the collected dataset using Arabic-BERT tokenizer and trained Arabic-BERT-base for additional 1 million steps using only Sudanese sentences.
- The pre-training of SudaBERT-base was carried out with a batch size of 32, and a learning rate of 1e-5 -an order of magnitude less than if the pre-training was from scratch-.
- We repeated the same approach to train SudaBERT-large, however this time we started the training from Arabic-BERT-large.
- Both SudaBERT-base and SudaBERT-large were pre-trained on a V3-8 TPU from Google Cloud Platform (GCP).

Evaluation Datasets

We evaluated our models on two NLU tasks: Sentiment Analysis (SA) and Name Entity Recognition (NER). Two of the SA datasets were written in Sudanese dialect, while the remaining datasets were written in Modern Standard Arabic (MSA) and other dialects. The following table shows the sizes of the datasets.

Dataset	Task	# of examples
AJGT	SA	1,800
ArSenTD-LEV	SA	4,000
ASTD	SA	3,188
LABR	SA	63,000
ANERCorp	NER	150,000*
Sudanese 1	SA	4,712
Sudanese 2	SA	2,116

Table 1: Sizes of different evaluation datasets. * the size of the NER dataset is measured by number of tokens.

Results

Dataset	Arabic-BERT		SudaBERT	
	Base	Large	Base	Large
AJGT	<u>91.6</u>	**93.5**	89.1	**93.5**
ArSenTD-LEV	<u>52.7</u>	**53.5**	50	53.2
ASTD	<u>59</u>	**60.7**	51.7	60.1
LABR	<u>83.5</u>	**85.7**	80.8	84.8
ANERCorp	<u>72.9</u>	**79**	70.2	78.4
Sudanese 1	57.7	59.3	<u>60.5</u>	**62.6**
Sudanese 2	76	79.4	<u>81.1</u>	**82**

Table 2: F1 score for models in each dataset. **Bold** is the best result for each dataset. <u>Underline</u> indicates the best base size model.

- Both Arabic-BERT-large and SudaBERT-large are generally better than their smaller counterparts.
- For both model sizes, SudaBERT outperformed Arabic-BERT on Sudanese dialect datasets, while Arabic-BERT performed better on other datasets.

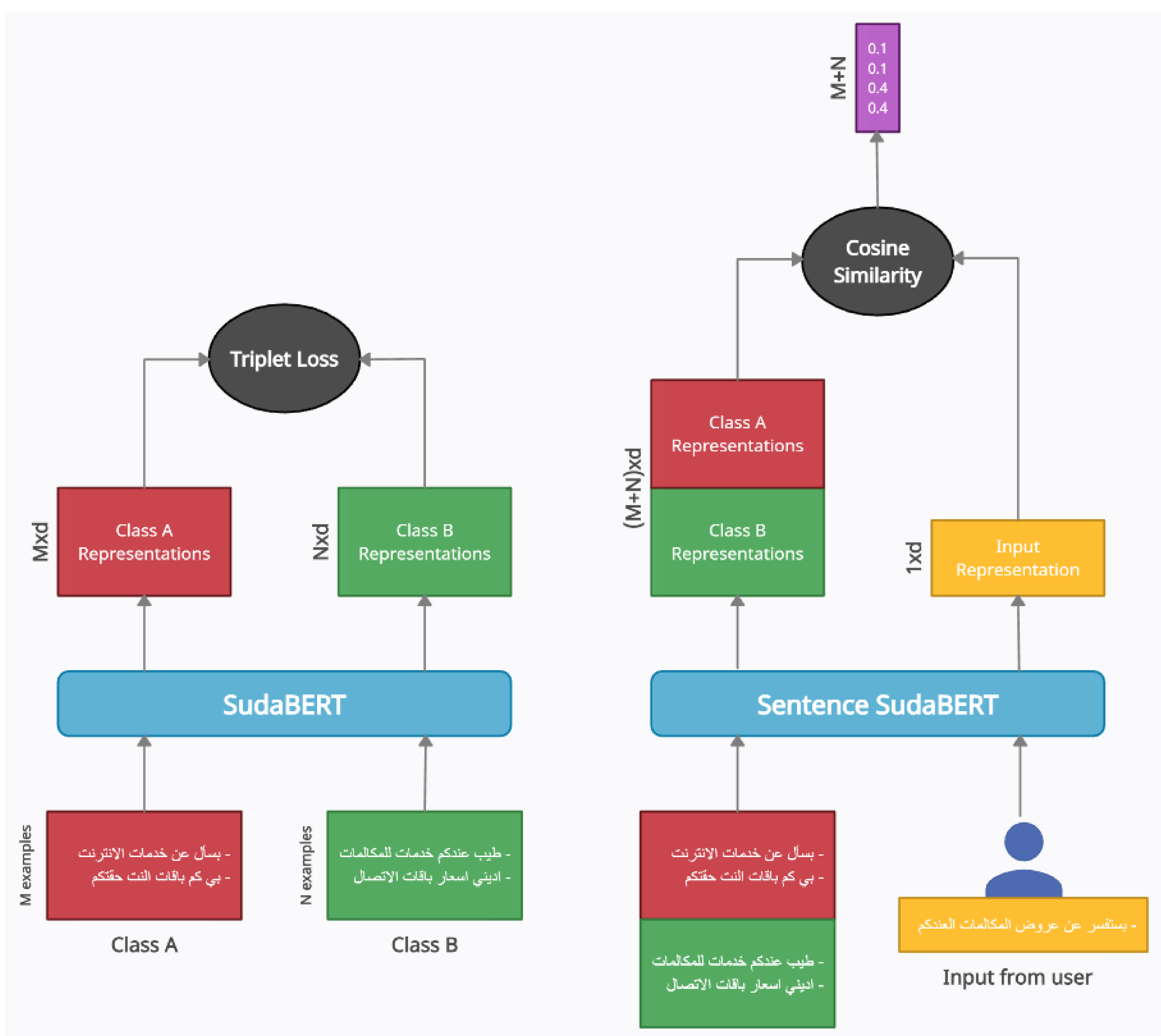

Figure 1: How we use SudaBERT in production.