NATURAL LANGUAGE PROCESSING 2022/23

Text Classification

MASSIVE: A 1M-Example Multilingual Natural Language Understanding Dataset with 51 Typologically-Diverse Languages

Margarida Vieira
up201907907

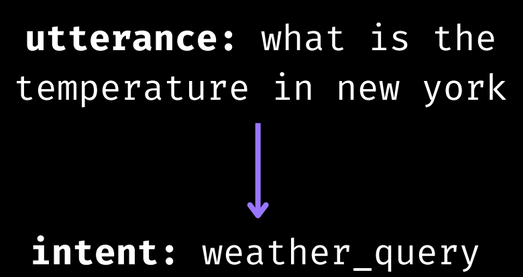
Nuno Costa
up201906272

Tiago Rodrigues
up201907021

Multiclass Classification Problem

This task's goal is to be able to **classify the intent of an utterance** - Intelligent Voice Assistant single-shot interactions.

As such, we'll be solving a **Multiclass Classification Problem**.



51 languages

60 intents

55 slot types

18 domains

MASSIVE | a multilingual dataset

A **1M-example dataset** composed of realistic, **human-created virtual assistant utterance text**, mainly in the interrogative and imperative form.

It has been translated by professional translators and it spans 51 different languages.

Results

Model	Multinomial NB			XGBoost Classifier			Logistic Regression		
	High	Low	Avg	High	Low	Avg	High	Low	Avg
TF-IDF	0.799 **ru-RU**	0.403 **my-MM**	0.739	0.789 **da-DK**	0.433 **my-MM**	0.752	0.817 **id-ID**	0.396 **my-MM**	0.760
Bag of Words	0.720 **pl-PL**	0.366 **my-MM**	0.655	0.782 **az-AZ**	0.429 **my-MM**	0.743	0.773 **en-US**	0.370 **my-MM**	0.711

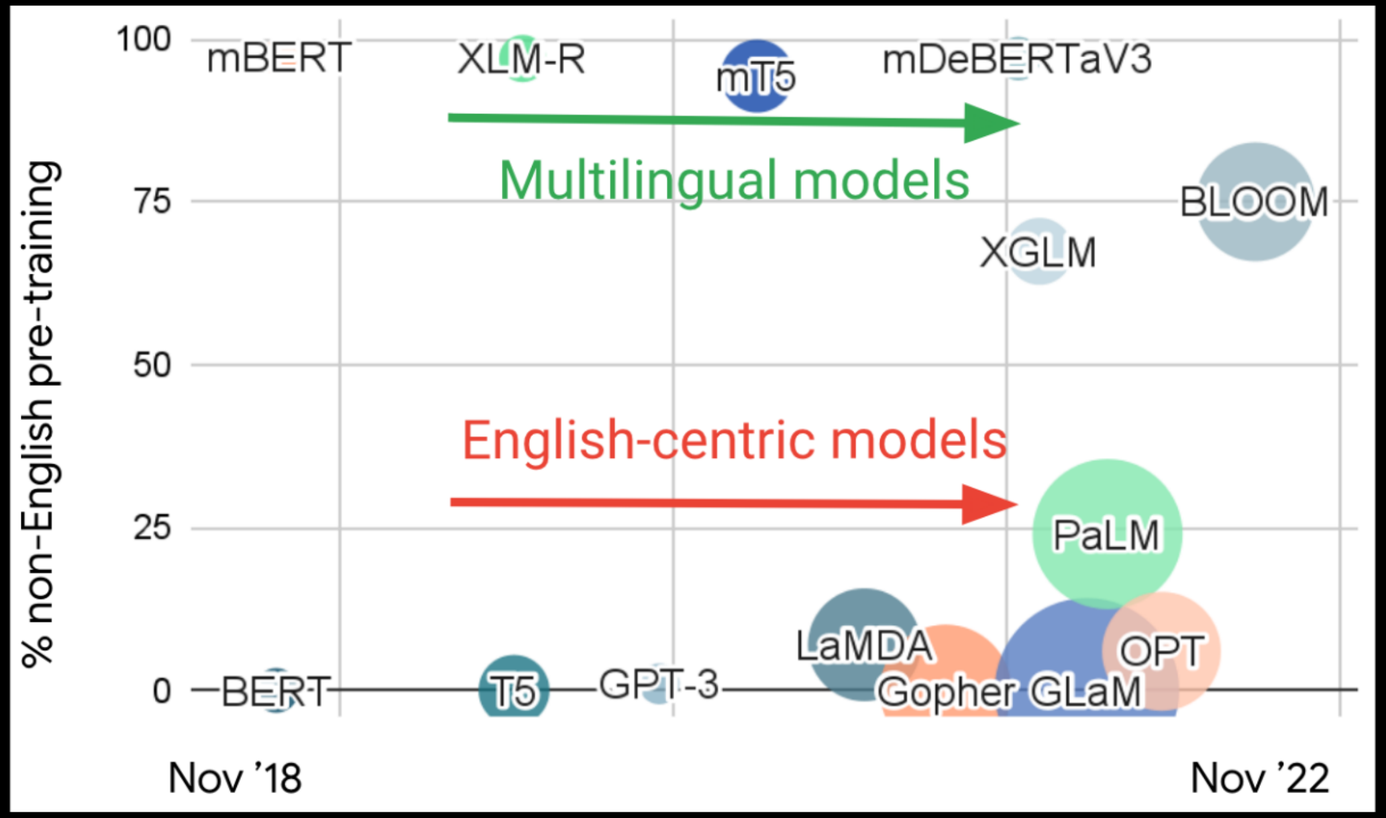
	High	Low	Avg
Our Baseline	0.745 **pl-PL**	0.119 **ja-JP**	0.638
XLM-R Base	0.883 **en-US**	0.772 **km-KH**	0.851

Used **Randomized Search** for *hyperparameter tuning*

Results are shown for the **best hyperparameters** with *lowercasing*, *tokenization*, and *stemming*

Multilingual Models

State-of-the-art



Multilingual Models

State-of-the-art

XLM-RoBERTa (encoder)

multilingual version of RoBERTa,

pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages

mDeBERTa (encoder)

pre-trained with the CC100 multilingual dataset, which comprises monolingual data for 100+ languages

mBart (encoder-decoder)

sequence-to-sequence multilingual encoder-decoder model

primarily intended for translation tasks

unfeasible to use due to size

BLOOM (decoder)

largest multilingual open-source model to date

autoregressive LLM with 176B parameters

pre-trained on and able to generate text in 46 natural languages and 13 programming languages

unfeasible to use due to size

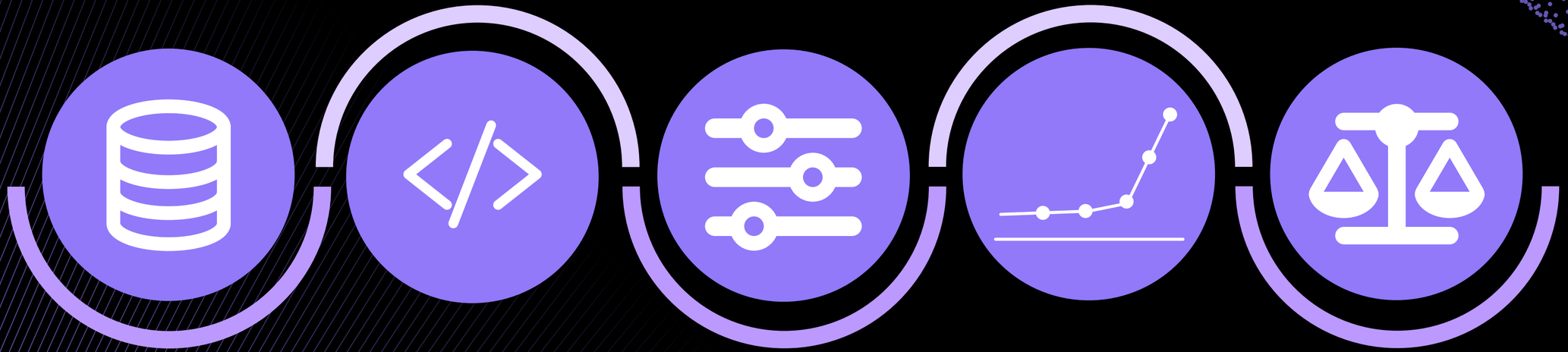
mT5 (encoder-decoder)

text-to-text transformer

pre-trained on a new Common Crawl-based mC4 dataset, covering 101 languages

unusual for classification tasks, since it is trained to generate the literal text of the class label instead of a class index

Our Approach



Small data preprocessing

punctuation removal
lowercasing
label encoding

Tokenization

using the model specific tokenizer, taking advantage of AutoTokenizer from the **transformers** library

Hyperparameter Definition

usage of **AdamW Optimizer** with:
learning rate = 2e-5
ε = 1e-8
and running with a **batch size** = 32

Model Training

ran for **2 epochs** for most models using a **CrossEntropy loss** function, taking advantage of **DataParallelism** (T4x2 GPU)

Model Evaluation

measure the model's **accuracy values** for the whole test set and also for each particular language to compare with the Paper's results

Results

XLM-RoBERTa

xlm-roberta-base

	High	Low	Avg
xlm-roberta-base Our Model	0.875 **en-US**	0.564 **km-KH**	0.801
xlm-roberta-base Paper's Model	0.883 **en-US**	0.772 **km-KH**	0.851

	Train Loss	Val. Loss	Val. Accuracy	Train Time	Val. Time
Epoch 1	0.861	0.674	0.832	3:22:03	0:11:30
Epoch 2	0.374	0.648	0.841	3:21:31	0:11:19

mDeBERTa

microsoft/mdeberta-v3-base

	High	Low	Avg
mdeberta-v3-base	0.880 **en-US**	0.508 **km-KH**	0.803
xlm-roberta-base Paper's Model	0.883 **en-US**	0.772 **km-KH**	0.851

	Train Loss	Val. Loss	Val. Accuracy	Train Time	Val. Time
Epoch 1	0.837	0.721	0.830	3:48:51	0:13:21
Epoch 2	0.353	0.714	0.844	3:49:01	0:13:28

Results

DistilBERT

distilbert-base-multilingual-cased

	High	Low	Avg
distilbert-base-multilingual-cased	0.835 **en-US**	0.030 **am-ET**	0.756
xlm-roberta-base Paper's Model	0.883 **en-US**	0.772 **km-KH**	0.851

	Train Loss	Val. Loss	Val. Accuracy	Train Time	Val. Time
Epoch 1	1.221	0.850	0.775	1:37:05	0:05:06
Epoch 2	0.631	0.798	0.796	1:36:58	0:05:09
Epoch 3	0.475	0.796	0.803	1:38:02	0:05:11

InfoXLM

microsoft/infoxlm-base

	High	Low	Avg
microsoft/infoxlm-base	0.871 **en-US**	0.514 **km-KH**	0.796
xlm-roberta-base Paper's Model	0.883 **en-US**	0.772 **km-KH**	0.851

	Train Loss	Val. Loss	Val. Accuracy	Train Time	Val. Time
Epoch 1	0.854	0.786	0.803	3:14:51	0:12:52
Epoch 2	0.457	0.736	0.828	3:15:03	0:12:47

Conclusions

Hard to match the baseline results for the Paper given their significant processing power advantage to train XLM-Roberta especially when it comes to hyperparameter tuning and the number of epochs the model is trained for

---

Processing the text might affect the model's ability to accurately predict results. Besides this, by encoding the labels we may loose some helpful context that the model could use for guidance

---

The results clearly showcase the strength of Transformers when compared with more traditional and simpler models on more complex tasks such as multilingual and multilabel classification