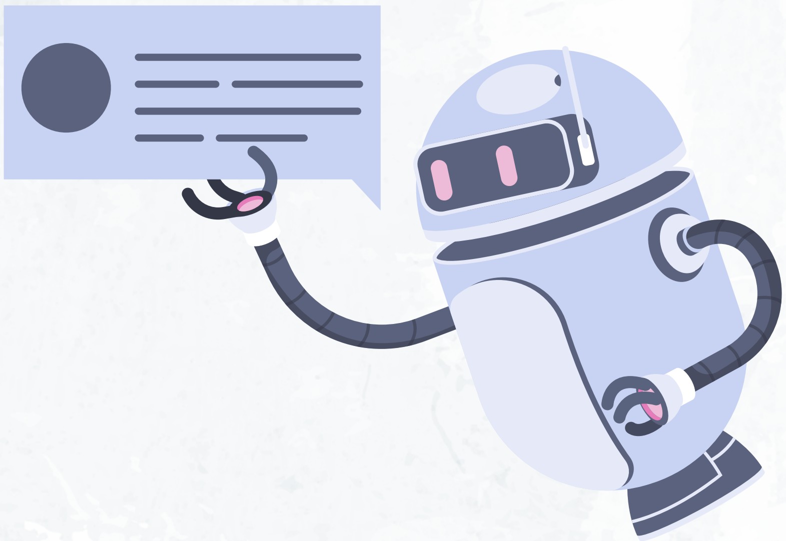
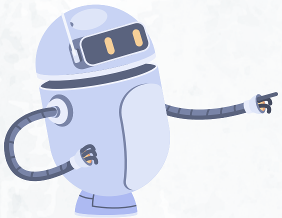
Multilingual Named Entity Recognition using MultiCoNER-II

School of Computing, Engineering & Intelligent Systems

Dharmesh Padvi

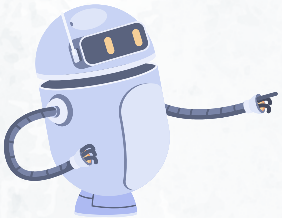




Introduction

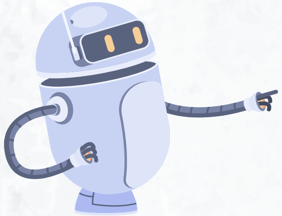
Named entity recognition(NER) is a sub-task in the field of **NLP**(Natural Language Processing). To be more specific NER is the text classification task in which a piece of the text is classified as their corresponding categories(person names, organisation names, etc).Basically, it involves two core tasks: **tokenizing** and **labelling** a piece of the information written in the human-readable languages. This field poses some challenges when it comes to recognizing the entities in different domains, for example, medical or creative names with more than one language.

Patrick PER and John Collison PER are two brothers from Limerick LOC .They became millionaires as teenagers when they sold their company, Auctomatic ORG , for $5 million.



Challenges In The Field Of NER

- ❑ **Complex named entities** such as(Creative names , Medical terminologies, etc.)
- ❑ **Multiple Domains**(Medical, Creative works, Organization names)
- ❑ **Multiple Languages**(English, Spanish, Hindi, Bangla, Chinese, Swedish, Farsi, French, Italian, Portuguese, Ukrainian, German)



The Aim of the project

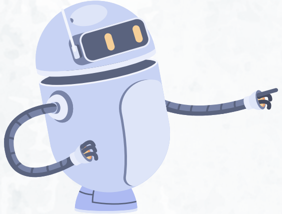
The aim of the project is to represent the development of a robust and high score achieving score ensemble **multilingual NER system** using pre-trained models such as RemBERT, mBERT, and xlm-r for 12 languages, and the weighted mean ensemble method to combine the predictions from these models 36 with multiple domains ; also, with the limited hardware support. To train these models, the MultiCoNER-2(12-monolingual subsets)dataset was used with the help of Hugging-face's transformers library and Pytorch library for deep learning.

HUGGING FACE

PyTorch

MultiCoNER

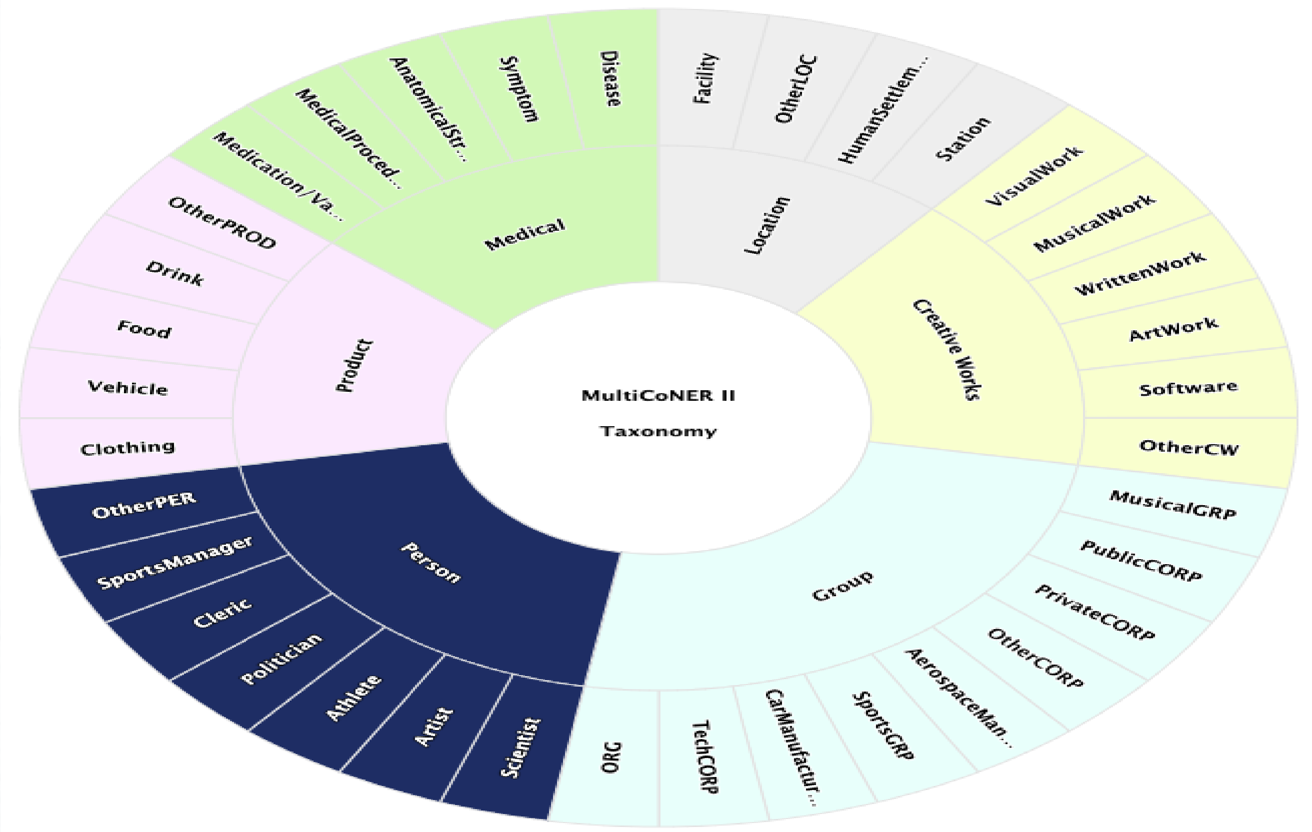
SemEval 2023



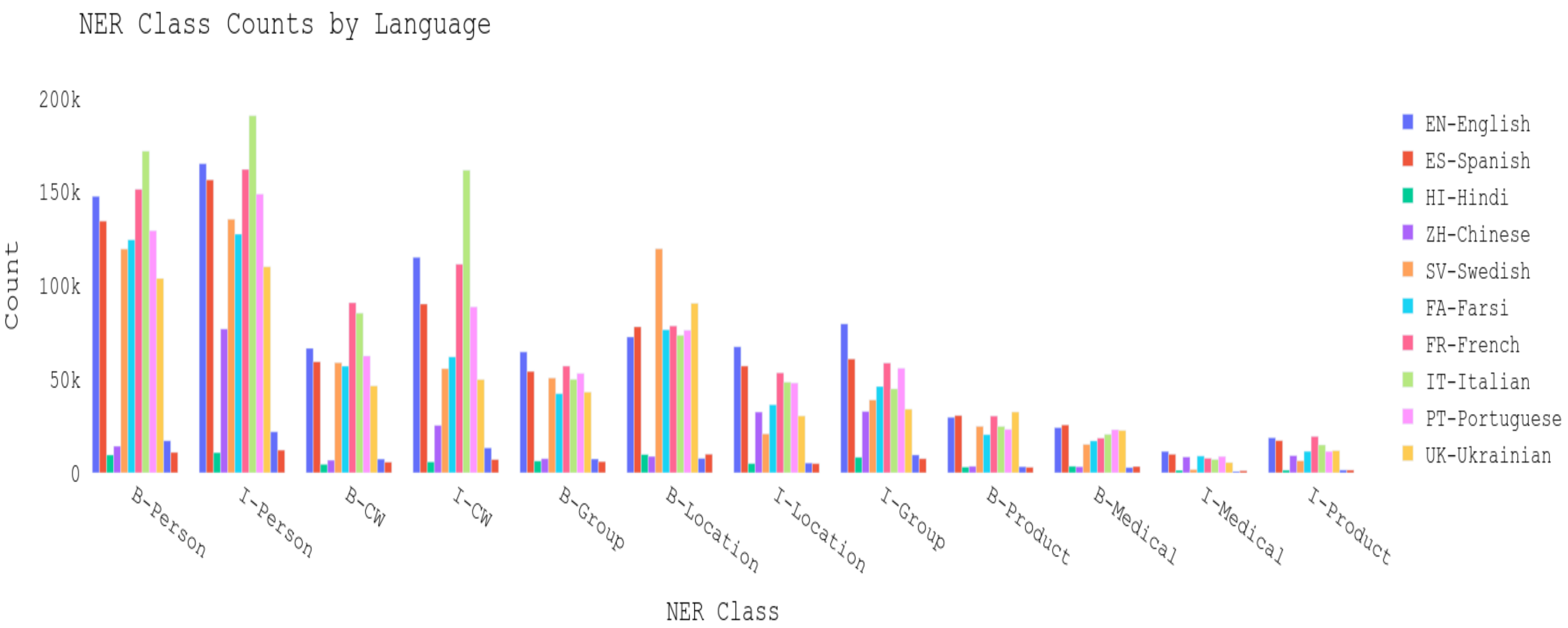
Database

MultiCoNER-II database consist of total 12 languages, such as English, Spanish, Hindi, Bangla, Chinese, Swedish, Farsi, French, Italian, Portuguese, Ukrainian, German, for fine-grained for Complex named entity recognition, also, its fine-grained taxonomy contains 36 classes, which represents the existing real world challenges for NER systems. The dataset is MultiCoNER-II is a second version(v2) of the dataset MultiCoNER released publicly as a part of as part of the *SemEval 2022 Task#11*[12]. Furthermore, It represents 3 domains such as wiki sentences(**LOWNER**) search queries(**ORCAS-NER**), questions(**MSQ-NER**), and. MultiCoNER taxonomy represents 6 ner-tagsets representing categories such as person, group, location, corporation, creative work. Table I. represents the detailed taxonomy with 6 ner-tagsets.

Database Taxonomy



Data Distribution



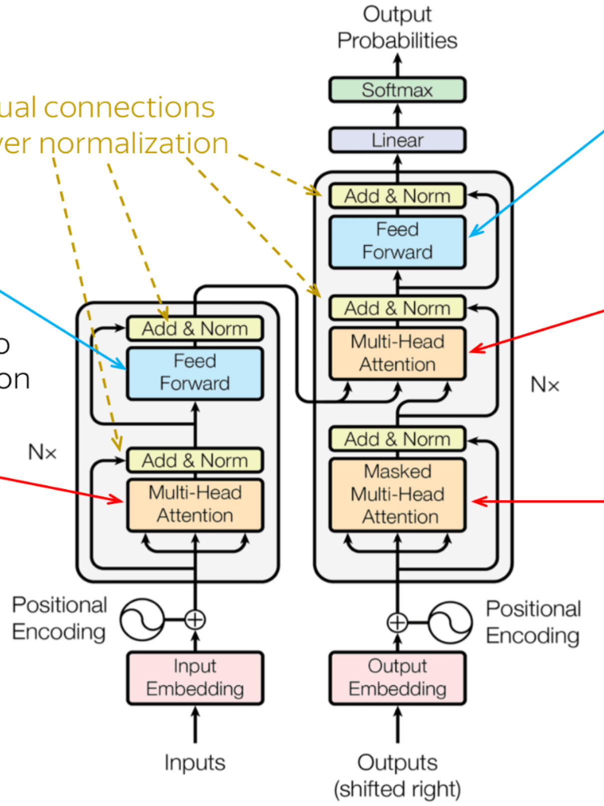
Transformer architecture

Residual connections and layer normalization

Feed-forward network: after taking information from other tokens, take a moment to think and process this information

Encoder self-attention: tokens look at each other

queries, keys, values are computed from encoder states



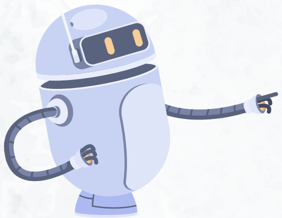
Feed-forward network: after taking information from other tokens, take a moment to think and process this information

Decoder-encoder attention: target token looks at the source

queries – from decoder states; keys and values from encoder states

Decoder self-attention (masked): tokens look at the previous tokens

queries, keys, values are computed from decoder states

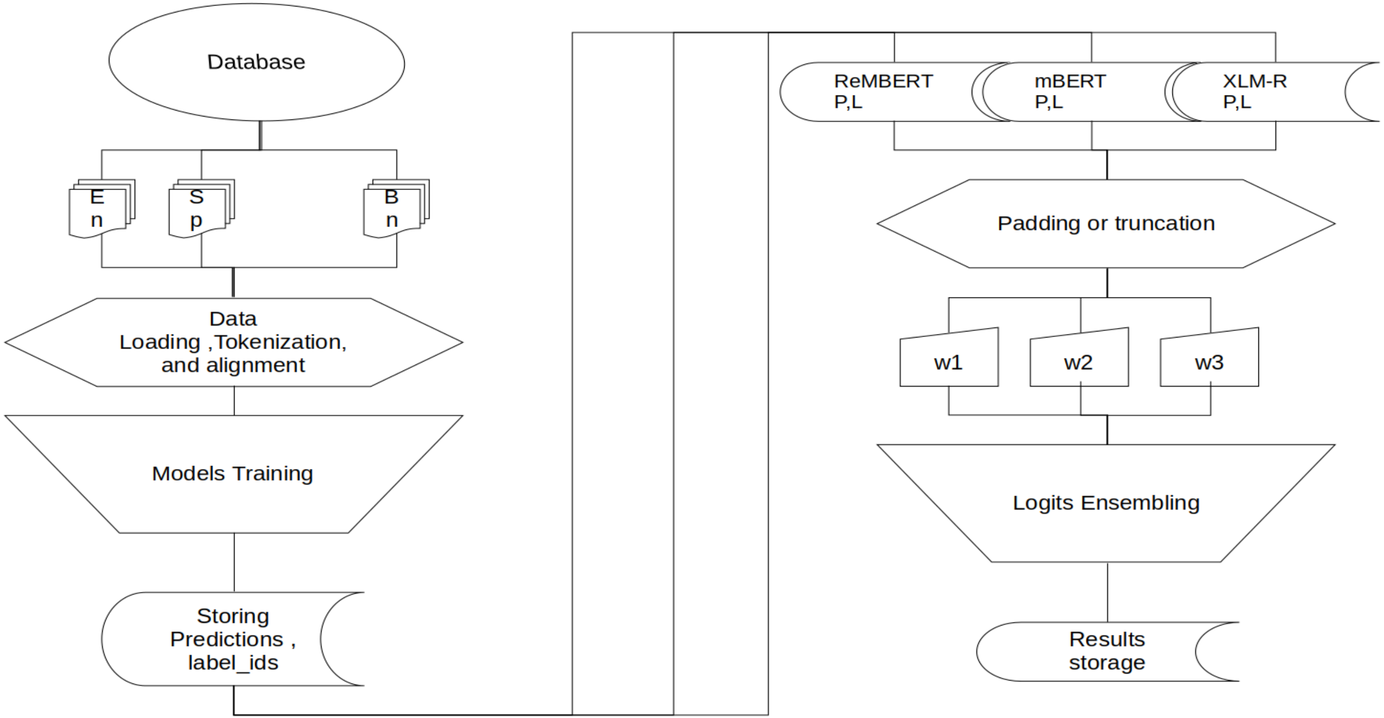


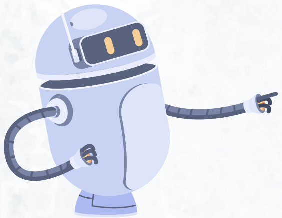
Model Selection

- ❑ **RemBERT** (Rebalanced mBERT)
- ❑ **mBERT** (Multilingual BERT where BERT stands for ‘Bidirectional Encoder Representations from Transformers’)
- ❑ **XLM-R** (xlm-roBERTa is a multilingual lingual version of the model RobBERTa where RoBERTa means optimised BERT pre-training approach)

High benchmarking achieving transformer based models like RemBERT, mBERT, xml-r had been used in the competition named **SemEval task-2 MultiCoNER-II**, and showcased the best predicting abilities of these transformers based multilingual models’

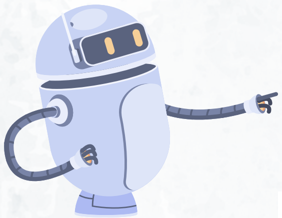
Project Implementation Process Diagram





Data Modeling & Fine-Tuning Parameters

We trained **36** models with **learning-rate 2.5e-5** for all languages, use **2 epochs** for all the languages, and models, excluding the RemBERT models for low-resource languages(Hindi, Chinese, German, and Bangla). To leverage maximum transfer learning from the pre-trained model, **we trained low-resource languages for 3 epochs**. Due to limitations of the hardware resources maximum token size of the Chinese and Bangla RemBERT models, we trained the model with **8 batch sizes**. For the rest of the RemBERT models, including mBERT, xlm-r, we use 16 batch-size. Batch size is also a hyperparameter which is associated with better model generalisation, and faster convergence. With **0.01 weight decay**, **100 warm up steps**, and **liner learning-rate_scheduler**, all 36 models were trained.



Algorithm for Ensemble Modeling

1. Initialize **max_seq_**lengths as an empty dictionary.

2. For each language in **languages**:(loop)

   Calculate the maximum sequence length across all models for that language and store it in **max_seq_lengths**.

3. Update **max_sequence_length** to be the maximum of all maximum sequence lengths.

4. Define the **weightage** for each model (Note: the total weight should be 1).

5. Initializing an empty list **ensemble_logits** to store the **ensembled logits.**

6. For each language in **languages**:(loop)

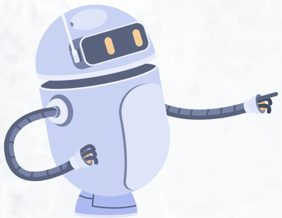
   Get the logits for each model.

   Ensure that all logits have the same shape (padding or truncating).

   Calculate the weighted average of logits for each example.

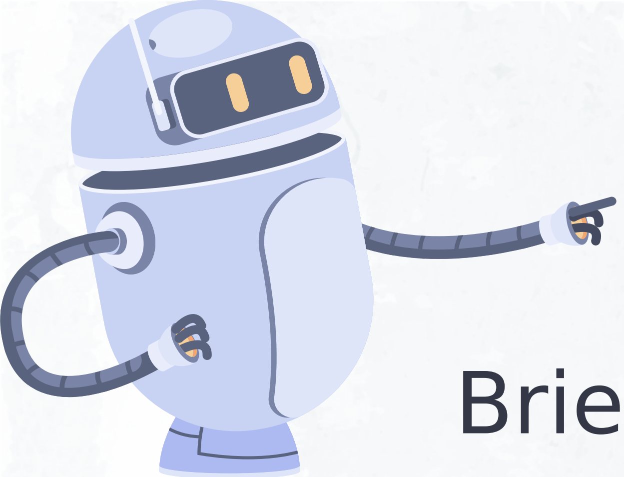
   Append the **ensembled logits** to the list.

7. Convert the **ensemble_logits list** to a NumPy array.

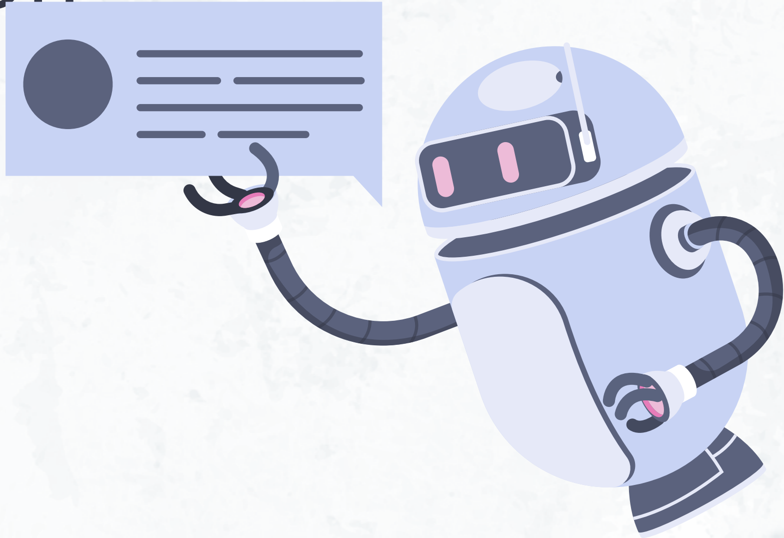


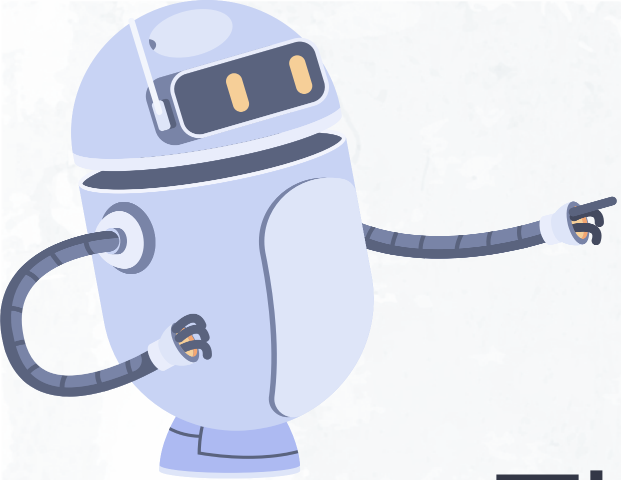
Results

Languages	RemBERT micro avg f1 scores	mBERT micro avg f1 scores	xlm-r micro avg f1 scores	Ensemble micro avg f1 scores
EN-English	0.95	0.91	0.89	0.91
ES-Spanish	0.96	0.92	0.9	**0.99**
HI-Hindi	0.99	0.88	0.87	**0.99**
ZH-Chinese	0.95	0.87	0.77	0.98
SV-Swedish	0.97	**0.95**	**0.94**	0.98
FA-Farsi	0.92	0.86	0.84	0.98
FR-French	0.95	0.92	0.9	0.98
IT-Italian	0.97	0.94	0.93	0.98
PT-Portuguese	0.96	0.93	0.92	0.98
UK-Ukrainian	0.96	0.92	0.92	0.98
DE-German	0.99	0.9	0.88	0.99
BN-Bangla	0.99	0.87	0.8	**1**



Brief Code Will Be Demonstrated On Google Colab





Thanks!