MULTILINGUAL AND CROSSLINGUAL FACT-CHECKED CLAIM RETRIEVAL

- INSTRUCTOR: ASUTOSH MODI
- GROUP: 4

PROBLEM STATEMENT

- Our objective is to develop a system that can accurately predict the most relevant facts for given social media posts across various languages.
- The system should be capable of handling both monolingual and cross-lingual retrieval scenarios.
- The system’s effectiveness will be measured through Mean Reciprocal Rank (MRR) and Precision@K.

LITERATURE SURVEY
X-FACT: A MULTILINGUAL FACT-CHECKING DATASET

- **Problem Statement:**
- Create a large multilingual dataset for fact-checking to improve systems beyond English.
- **Dataset Overview:**
- **X-Fact** dataset includes **31,189 claims** in **25 languages** across **11 language families.**
- Split into 3 parts:
  - **Training**: 19,079 claims (13 languages)
  - **Development**: 2,535 claims (12 languages)
  - **Testing**: 3 test sets:
    - **In-domain Test Set (α1)**: Similar to training data.
    - **Out-of-domain Test Set (α2)**: Same languages, different sources.
    - **Zero-shot Test Set (α3)**: New languages not in training set.
    -

Source: X-Fact: A New Benchmark Dataset for Multilingual Fact Checking (aclanthology.org)

X-FACT: A MULTILINGUAL FACT-CHECKING DATASET

- **Modeling Approaches:**
- **mBERT-based models** (multilingual BERT for 104 languages).
  - **Claim-Only Model**: Predicts true/false based on claim text only.
  - **Attention-Based Evidence Aggregator (Attn-EA)**: Uses web snippets as supporting evidence.
  - **Models with Metadata (+Meta)**: Adds contextual information like language, website name, and date.
- **Results:**
- **Claim-Only Model**: F1 score ~38.2% on in-domain claims.
- **Attn-EA**: Small improvement, F1 score ~41.9%.
- **Limitations:** Performance dropped significantly on the out-of-domain and zero-shot test sets. This means that models struggled to generalize to claims from new sources or new languages. For example, the best-performing model on the zero-shot test set had an F1 score of only 16%.

DATASET ANALYSIS AND PREPROCESSING

- **Dataset Overview:** The dataset consists of three main files:
- **Fact Checks (`fact_checks.csv`)**
- **Fields**:
  - `fact_check_id`: Unique identifier for each fact check.
  - `claim`: Original claim, its translated version, and the language.
  - `instances`: List of timestamps and URLs where the fact check is mentioned.
  - `title`: Original title, its translated version, and the language.

fact_check_id	claim	instances	title
12	('!! Ø¨ÙŠØ´Ø±Ø¨ Ù…ÙŠØ§Ù‡ Ù…Ø¹Ø¯Ù†ÙŠØ© Ù.	[(None, 'https://dabegad.c	('Ø-Ù,ÙŠÙ,Ø©Ø´Ø±Ø¨ Ø§Ù„Ù³Ù
13	('" As vacinas nÃ£o passaram pelos protocolos de te	[(1614511874.0, 'https://c	('Fact Check. As vacinas contra a

- **Social Media Posts (`posts.csv`)**
- **Fields**:
  - `post_id`: Unique identifier for each social media post.
  - `instances`: List of timestamps and social media platforms where the post appears.
  - `ocr`: Text extracted from images in the post, along with translations.
  - `verdicts`: Labels attached by platforms (e.g., "False information").
  - `text`: The main text of the post and its translated version.
  - **posts**

post_id	instances	ocr	verdicts	text
2751	[(1604555943.0, 'fb')]	[('t YEARNING- HOME F(	['False information']	('!Boletas tiradas!
2752	[(1604588844.0, 'fb')]	[('UNITED STATES POST	['False information']	('!Boletas tiradas!

- **Mapping (`pairs.csv`)**
- **Fields:**
  - **`fact_check_id`: Links to `fact_checks.csv`.**
  - **`post_id`: Links to `posts.csv`.**

post_id	fact_check_id
2228	33
2228	23568

- **Format of Predictions:**
- A single JSON file where each key is a `post_id` (as a string) and the corresponding value is a list of up to 10 `fact_check_id`s (as integers) that are most relevant to that post.

- **Example :**

```
{
  "0": [5, 10, 2, 65, 15, 255, 11, 8, 420, 502],
  "1": [12, 13, 5, 0, 125, 450, 220, 18, 49, 51],
  "2": [444, 4, 7, 18, 29, 55, 263, 178, 99, 82],
  ...
  "3840": [11, 3, 507, 624, 177, 39, 20, 66, 344, 327]
}
```

- **Key:** `post_id` (as a string).
- **Value:** A list of 10 `fact_check_id`s (as integers) that your system predicts to be the most relevant fact checks for that post.

CHECK POSTS PER VERDICT

1. False information 12408
2. Partly false information 3410
3. Missing context 1355
4. False information. 1147
5. Altered photo 493
6. Partly false information. 351
7. Partly False 271
8. Missing context. 256
9. False 241
10. False information and graphic content 222
11. Altered video 108
12. Missing Context 88
13. Sensitive content 33
14. Altered photo/video. 15
15. Altered Photo/Video 14
16. False headline 2
17. Support your streamers by sending them stars 2
18. Altered photo/video 1

MODEL

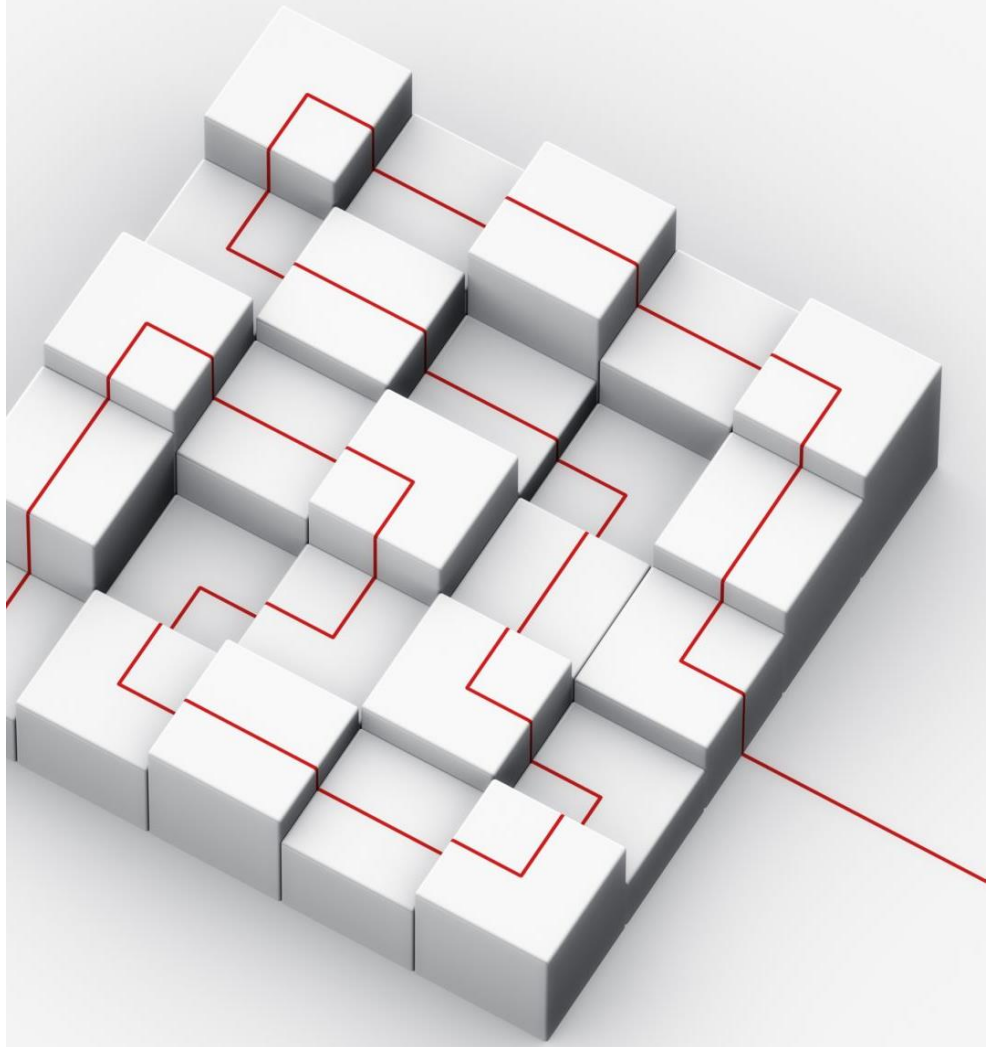
Motivation: Based on X-FACT research, we use a **Claim-Only Model** with **mBERT (Multilingual BERT)** to classify whether a fact-check claim is relevant to a social media post.

Model Architecture:

- **Input**: Paired social media post and fact-check claim, tokenized using mBERT.
- **Output**: Binary label (1 = relevant, 0 = irrelevant).
- **mBERT**: Pre-trained multilingual model utilizing attention and positional encodings to link claims and posts across languages.

Tokenization & Data Formatting:

- **Paired Inputs**: Tokenized post and claim using Huggingface tokenizer.
- **Features**: Inputs include input_ids, token_type_ids, labels, and attention_mask.



MODEL ARCHITECTURE

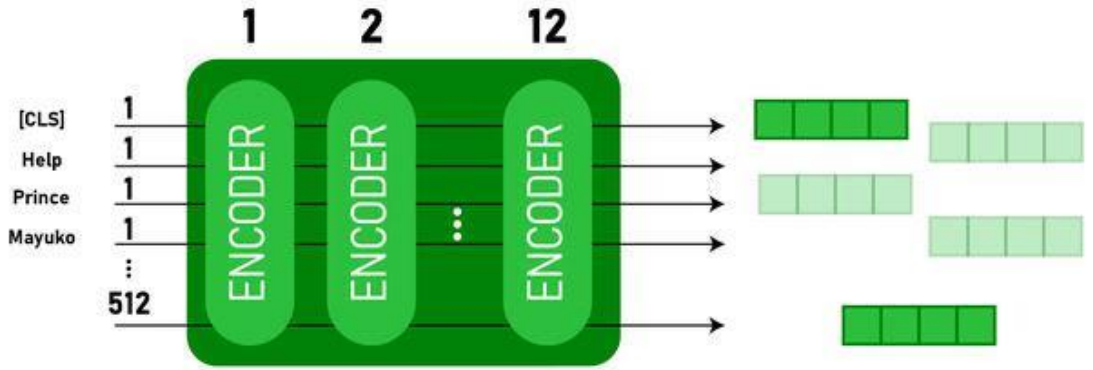
1. Introduction to mBERT (Multilingual BERT):

- **What is mBERT?** Multilingual BERT (mBERT) is a version of BERT (Bidirectional Encoder Representations from Transformers) that supports multiple languages. It's pre-trained on the top 104 languages with Wikipedia. This architecture enables cross-lingual NLP tasks, making it highly applicable for multilingual fact-checking.
- ***Example: "The bank is situated on the _____ of the river."***
- *In a unidirectional model, the understanding of the blank would heavily depend on the preceding words, and the model might struggle to discern whether "bank" refers to a financial institution or the side of the river.*
- *BERT, being bidirectional, simultaneously considers both the left ("The bank is situated on the") and right context ("of the river"), enabling a more nuanced understanding. It comprehends that the missing word is likely related to the geographical location of the bank, demonstrating the contextual richness that the bidirectional approach brings.*
-

source:GFG

- **Why mBERT for Fact-Checked Claim Retrieval?** Fact-checking on social media involves posts from multiple languages and regions. mBERT can handle cross-lingual retrieval, allowing the system to retrieve similar claims from a multilingual dataset, and thus facilitate fact-checking.

- **2. mBERT Architecture Overview:**
- **The Transformer Architecture:**
  - BERT is based on the Transformer architecture, which uses an **Encoder-Decoder model**. In mBERT, only the **encoder** is used.
  - mBERT follows a **bidirectional** approach, where the model takes into account both left and right contexts for all tokens, ensuring better understanding.



- **Input Representation:**
  - WordPiece tokenization: mBERT uses **WordPiece tokenization** to handle multilingual data.
  - Special tokens [CLS] and [SEP] are used for sentence classification and separation of input sequences.

3. mBERT for Fact-Check Claim Retrieval:

- **Claim Matching Process**:
  - For fact-checked claim retrieval, social media posts can be tokenized and passed through mBERT to generate embeddings. These embeddings are then compared with fact-checked claim embeddings to determine similarity.
- **Cross-lingual Retrieval**:
  - mBERT is fine-tuned on a multilingual claim dataset to allow **cross-lingual fact-check retrieval**. This enables the model to retrieve relevant fact-checked claims even when the source post is in one language and the fact-check database is in another.

4. Fact-Check Retrieval Pipeline Using mBERT:

- **Data Preprocessing**:
  - Tokenization and embedding generation for both social media posts and fact-checked claims.
- **Retrieval Mechanism**:
  - Cosine similarity is commonly used to compute the similarity between the embeddings of a new claim and fact-checked claims in the dataset.
- **Ranking and Filtering**:
  - Top-N most similar claims are retrieved, and thresholding is applied to discard irrelevant ones.

5. Challenges in Fact-Check Retrieval:

- **Handling Noisy Data**: Social media posts often contain noise (e.g., informal language, abbreviations). mBERT's WordPiece tokenization helps handle some of this, but further preprocessing steps like normalization might be needed.
- **Code-switching**: Social media users often switch between languages. mBERT can handle this due to its multilingual training.
- **Out-of-Vocabulary Words**: OOV words are often split into subwords, ensuring better handling by the model.

6. Fine-tuning mBERT for Fact-Check Retrieval:

- mBERT is fine-tuned on a fact-checking dataset using supervised learning. It is trained to minimize the loss on classification tasks, such as determining whether a claim matches a fact-checked claim or not.

- **Colab Files :**

- **DataPreprocessing :**
- https://colab.research.google.com/drive/12NV9eSujjwmwIIct3XKSCe1DJv0qEA0D?usp=drive_link
- **ClaimOnlyModel :**
- https://colab.research.google.com/drive/1l-u7iUIsk0Hf9HpC1LFh5GprzPVRZEw-?usp=drive_link
- **ContrastiveLoss Implementation:**
- https://colab.research.google.com/drive/1Lmhx1vubBI0nSz0CaYQz7B8pWXZdwD7h?usp=sharing
- **Merged_Columns_Input :** https://drive.google.com/file/d/12-5hZ8EYFPWyfzuCeEHOb2fi-8pxiFOt/view?usp=sharing

EXPERIMENT CONFIGURATION AND RESULTS

- **Experiment Configuration:**
  - **Training**: On paired input sequences of posts and claims.
  - **Evaluation Metrics**: Accuracy, precision, recall, and F1-score.

- **Results:**
  - Accuracy : ~85%
  - Precision : ~82%
  - Recall : ~84%
  - F1-Score : ~83%

Use Contrastive oloss function and select hard negative facts

LIMITATIONS AND FUTURE DIRECTION

- **Cross-lingual Limitations:** While mBERT handles many languages, its performance may degrade with languages that have less pre-training data (e.g., low-resource languages).

- **Future Direction** -
  - **Additional Features:** Including OCR as additional features could improve model performance.
  - **Handling Ambiguity:** Developing more sophisticated models that can better handle ambiguous or partial matches between posts and claims.
  - **Language wise prediction:** Use the information regarding language of claims and posts to do language wise retrieval of facts and check how the accuracy varies.