Welcome !

Agenda

- Introduction
- Problem Statement
- Motivation
- Methodology :How transcription work ！
- Methodology :How whisper model work and its components
- Why We Used the Whisper Model and Hugging Face Technology
- Methodology :Fine tune process
- Data Collection
- Discussion: Problems Faced and Positive Aspects
- Discussion: Challenges
- Conclusion: Future Plans

Introduction

North Africa's rich cultures and languages include the unique Darija dialect, spoken by millions and challenging for linguists due to its fluidity and lack of standardization. This project develops a robust model to extract text from Algerian Darija audio and video, capturing its nuances to bridge language barriers and enhance cultural understanding.

Problem Statement

- Turning spoken Darija into written text is challenging due to:
  - Informal structure.
  - Regional differences.
- Existing models struggle to capture Darija's nuances.
- Despite advancements, extracting text from Darija remains complex.
- Accurate speech extraction can:
  - Improve machine translation.
  - Connect North Africa with the world.
  - Make cultural resources more accessible.
- Our project addresses these issues by:
  - Collecting diverse Algerian Darija samples.
  - Fine-tuning the Whisper model.
- Goals:
  - Enhance human-computer interaction.
  - Foster cultural exchange and understanding.
  - Reliably recognize Darija speech for various applications.

Motivation

Cultural Relevance

Capturing and transcribing this language variety would enhance cultural understanding and communication.

User Engagement:

By offering accurate transcription of Algerian slang, we enhance user engagement and satisfaction, meeting the needs of speakers who prefer to communicate in their local dialect.

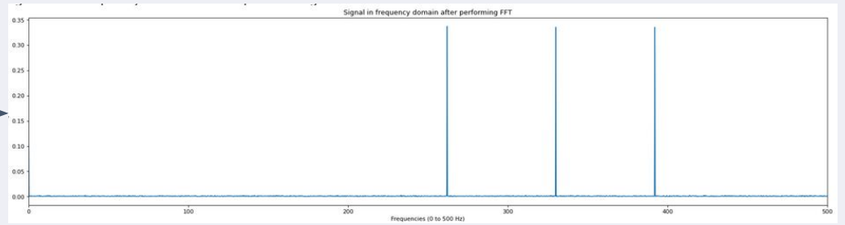
Support Language:

Despite its challenges, Darija is the main spoken language in North Africa, especially in Algeria. Algerian Darija uniquely blends Arabic (Oulad Nail tribe), Amazigh, French, Spanish, and Turkish, reflecting the region's rich cultural heritage.

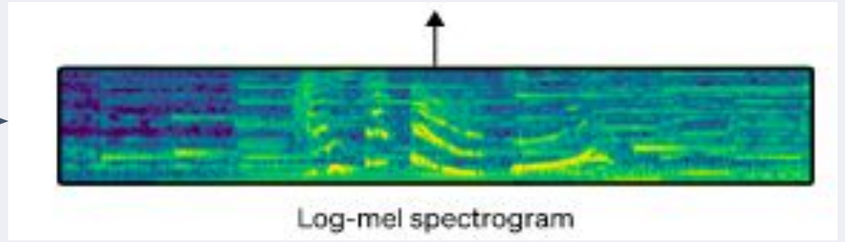
Methodology :How transcription work ！

1. Audio and sound

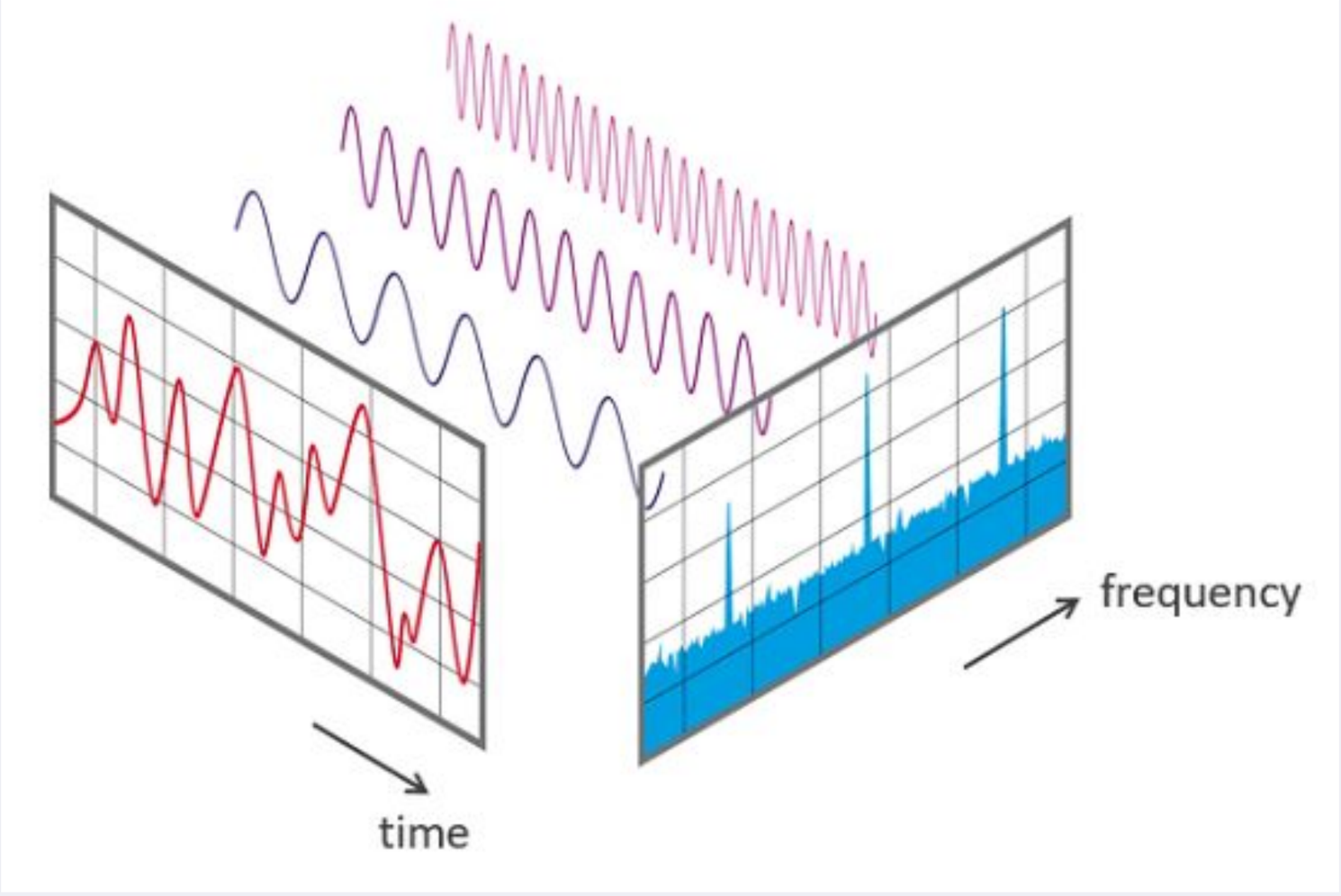
2. Frequencies



3. inputs to the model (“Features”)



4. Generate tokens (sub-words)

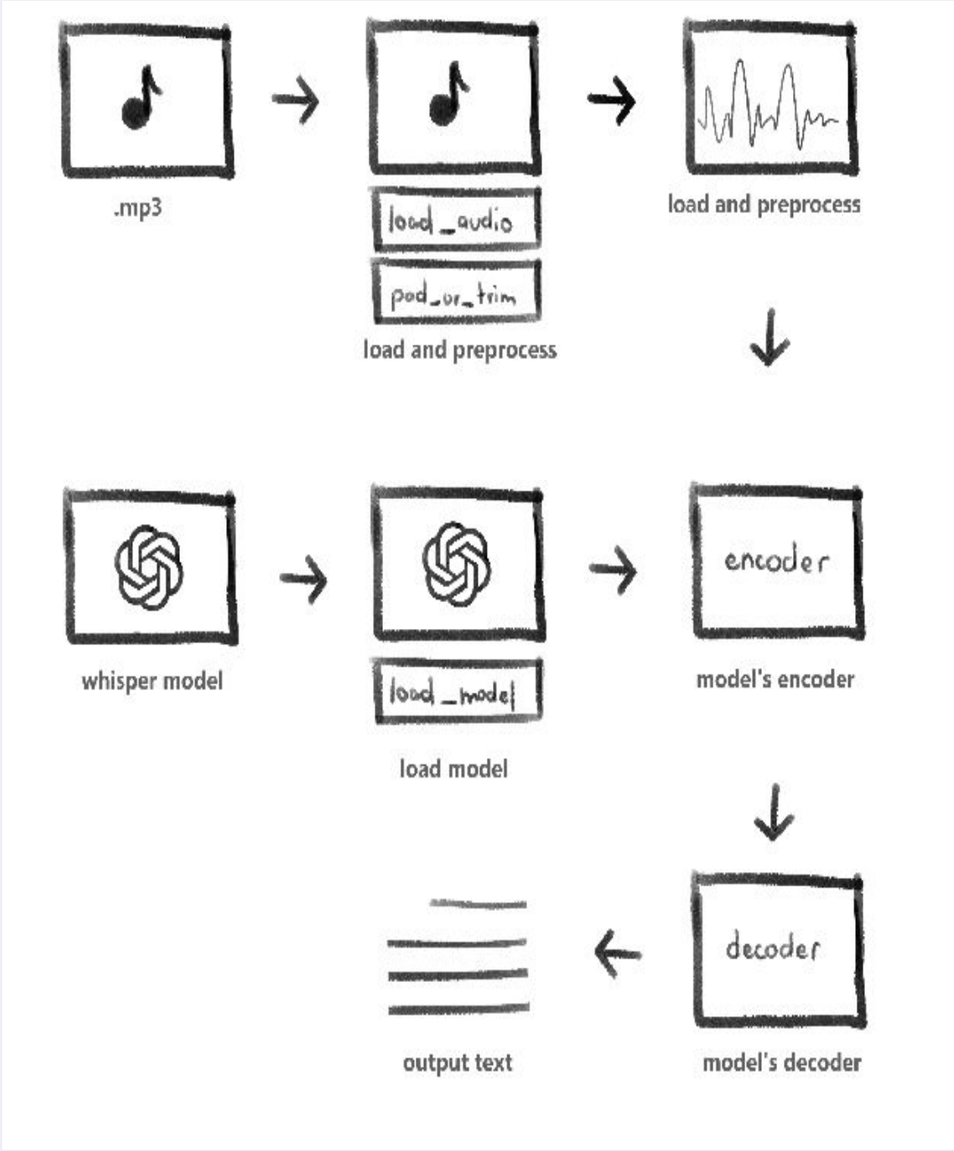


Fourier transform to decompose audio into series of frequencies

Methodology :How whisper model work

Whisper is a Transformer based encoder-decoder model, also referred to as a sequence-to-sequence model. It is a pre-trained model for **automatic speech recognition (ASR)** published in September 2022 by the authors Alec Radford et al. from OpenAI.

- Trained on more than 680,000 hours of labeled data this allows the model to **generalize well** to various accents, speaking styles, and background noise conditions, making it highly **robust and versatile** for real-world applications.
- transcribe speech in over **100 different languages**.
- Operates on a 30 seconds total window input



For more details read on : [fine_tune_whisper]

Methodology :How whisper model work

Whisper is a Transformer based encoder-decoder model, also referred to as a sequence-to-sequence model. It is a pre-trained model for **automatic speech recognition (ASR)** published in September 2022 by the authors Alec Radford et al. from OpenAI.

- Trained on more than 680,000 hours of labeled data this allows the model to **generalize well** to various accents, speaking styles, and background noise conditions, making it highly **robust and versatile** for real-world applications.
- transcribe speech in over **100 different languages**.
- Operates on a 30 seconds total window input

Size	Parameters	English-only model	Multilingual model
tiny	39 M	`tiny.en`	`tiny`
base	74 M	`base.en`	`base`
small	244 M	`small.en`	`small`
medium	769 M	`medium.en`	`medium`
large	1550 M	N/A	`large`

For more details read on : [fine_tune_whisper]

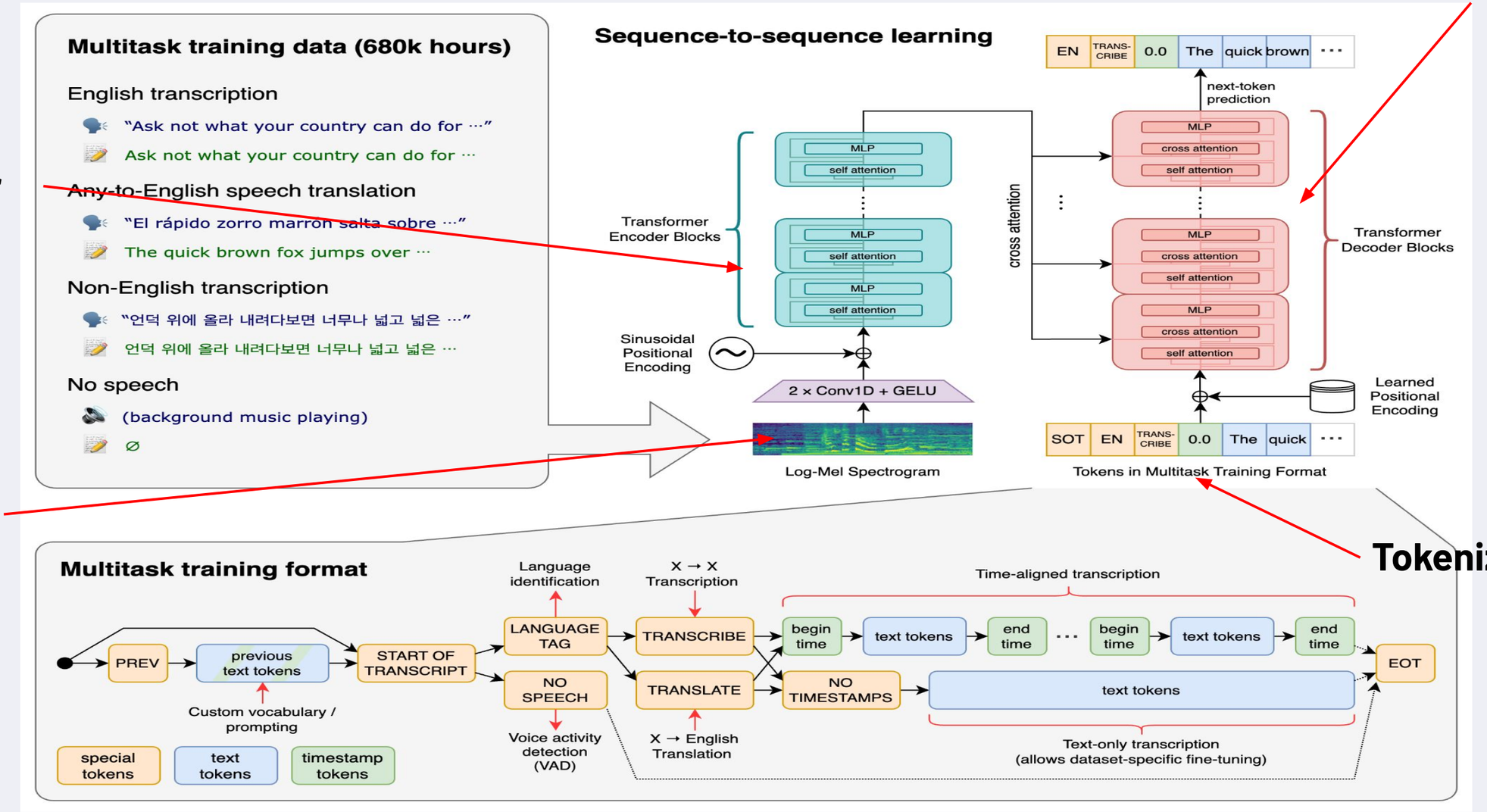
Methodology :How whisper model work

Decoder

Encoder

Input

Tokenizer



Methodology :whisper model components

Encoder

Converts the input audio signal into a **series of high-dimensional feature vectors** that capture the information of the speech signal. These feature vectors are then passed on to the **decoder** for further processing.

Tokenizer

Converts the **raw text output** from the **decoder** into a sequence of tokens, which are then converted into numerical representations suitable for further processing

Decoder :

Takes the **high-dimensional feature vectors** from the **encoder** and decodes them into text. It uses a recurrent neural network (RNN),to generate the **sequence** of characters corresponding to the input speech signal.

Methodology :whisper model components

sequence-to-sequence

consist of an **encoder and a decoder working together.** This architecture is commonly used in tasks like machine translation and speech-to-text.

Transformers

The Whisper model is known for its efficient and fast processing, able to generate accurate transcripts in real-time. This allows us to provide a seamless and responsive experience for our users.

Why We Used the Whisper Model

Audio Transcription

The Whisper model is a powerful tool for accurately transcribing audio recordings into text. Its **advanced natural language processing capabilities** allow it to handle a wide range of accents and dialects with high precision.

Multilingual Support

Whisper is a multilingual model, trained on data in hundreds of languages. This makes it an ideal choice for our project, which requires handling diverse audio inputs from various regions and backgrounds.

Efficient Performance

The Whisper model is known for its efficient and fast processing, able to generate accurate transcripts in real-time. This allows us to provide a seamless and responsive experience for our users.

Methodology :Fine tune process

Setup Environment	Load Dataset	Load Pretrained Model
Libraries needed : - transformers - accelerate - evaluate - Torch - tensorboard	From **hugging face** or read from disk	- Prepare Feature Extractor, Tokenizer and Data - Load WhisperFeatureExtractor - Load WhisperTokenizer - Combine To Create A WhisperProcessor - Load the model which performs the sequence-to-sequence

Fine tune whisper

Methodology :Fine tune process

Prepare Data for the model	Define the Training Arguments and evaluation	Training
- Load and resample the audio - Use the feature extractor to compute the log-Mel spectrogram input features from our 1-dimensional audio array. - encode the transcriptions to label ids through the use of the tokenizer.	- Use the Word Error Rate (WER) metric. - Pre-processes the raw audio-inputs. - The model which performs the sequence-to-sequence mapping. - Tokenizer	- Training will take approximately **5-10 hours** depending on your GPU

Fine tune whisper

Why We Use Hugging Face Technology

State-of-the-Art Models

Hugging Face provides access to cutting-edge language models like Whisper, which have been pre-trained on vast amounts of data to achieve exceptional performance on a variety of tasks.

Ease of Use

The Hugging Face ecosystem offers simple and intuitive APIs that make it easy for developers to integrate advanced AI capabilities into their applications without the need for complex model training.

Collaborative Development

Hugging Face has a vibrant community of AI researchers and engineers who continuously improve and expand the model repository, ensuring access to cutting-edge advancements.

Data collection

The data collection phase was critical for our speech-to-text project, aiming to capture a wide range of Arabic dialects, specifically Darija and Kabyle. Our goal was to compile a diverse and high-quality dataset that would ensure the accuracy and reliability of our model. This phase involved various strategies to gather extensive audio recordings and their corresponding transcriptions, forming the foundation for training and refining our speech-to-text system.

To collect the necessary data for our project, we employed a multi-step approach involving web scraping, a user-friendly form, and our own direct data collection efforts.

Data Collection

Initial Exploration

We started by researching existing speech-to-text tools for Arabic, particularly Darija. Subsequently, we utilized the YouTube API to collect videos featuring Darija and Kabyle speech, allowing us to gather a diverse set of audio samples from online sources.

This step involved evaluating various available technologies and their applicability to our specific needs.

Website dev and Manual Form Input

Initially, we thought about developing a simple web platform to streamline data collection, allowing users to upload audio files with transcriptions. To complement this, we created a straightforward form for manual input of audio data and transcriptions. Although these methods generated some positive responses, the overall data volume was insufficient for our project needs.

Data Collection

Manual Data Collection

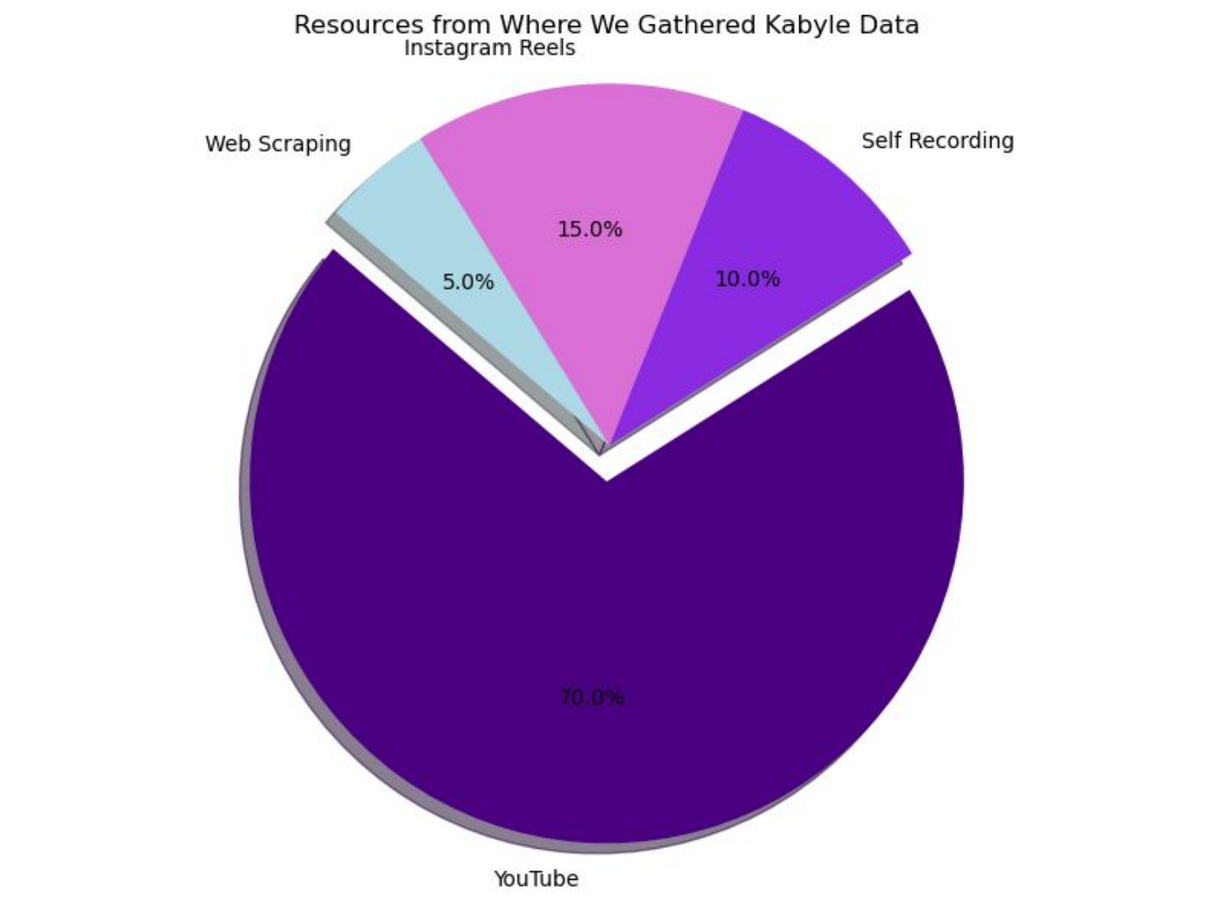
To enhance our dataset, we resorted to manual collection methods. We gathered additional audio data from various sources, including YouTube videos, Instagram reels, and personal voice recordings in Kabyle.

Data Compilation and Upload

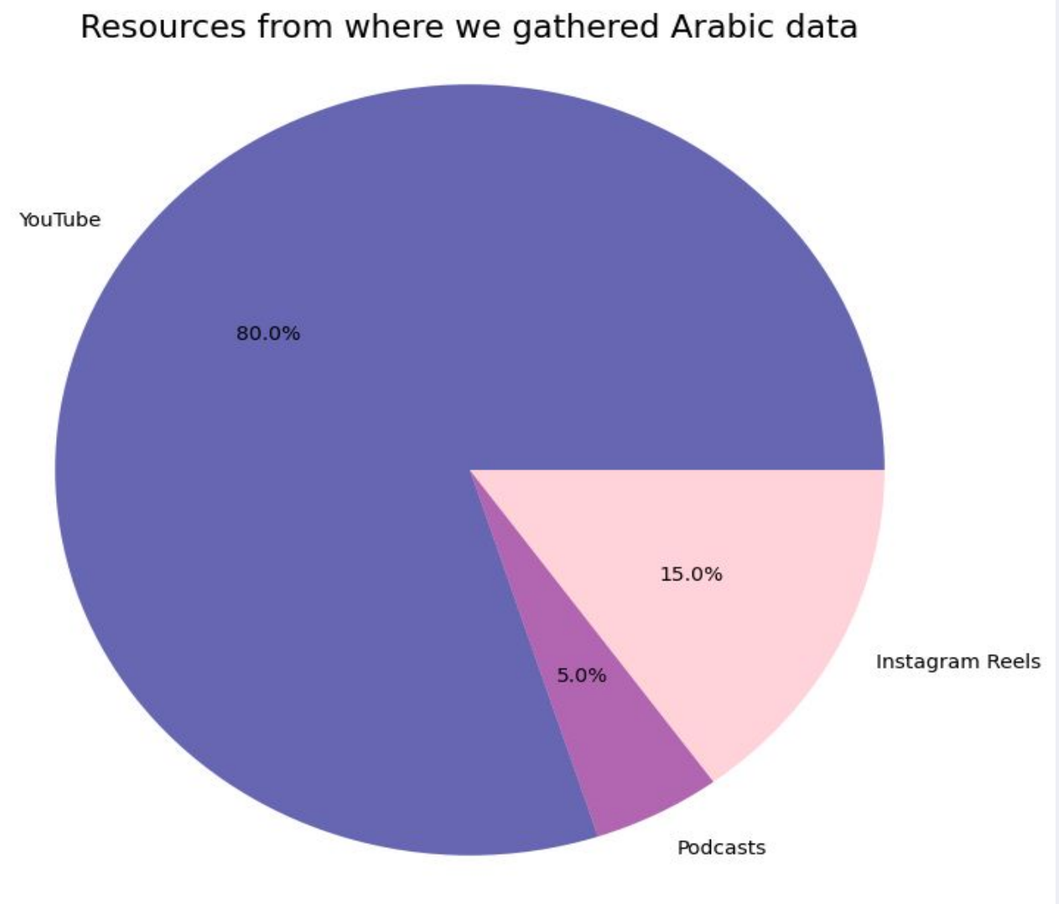
After gathering a substantial amount of data, we organized and uploaded it to Hugging Face, ensuring it was accessible for training our speech-to-text model.

Data Collection – some statistics

To get an overall idea of the different resources used in the manual data collection phase, we present pie charts for both the Kabyle and Darija (Arabic) datasets. These charts illustrate the various sources from which we gathered our audio data.

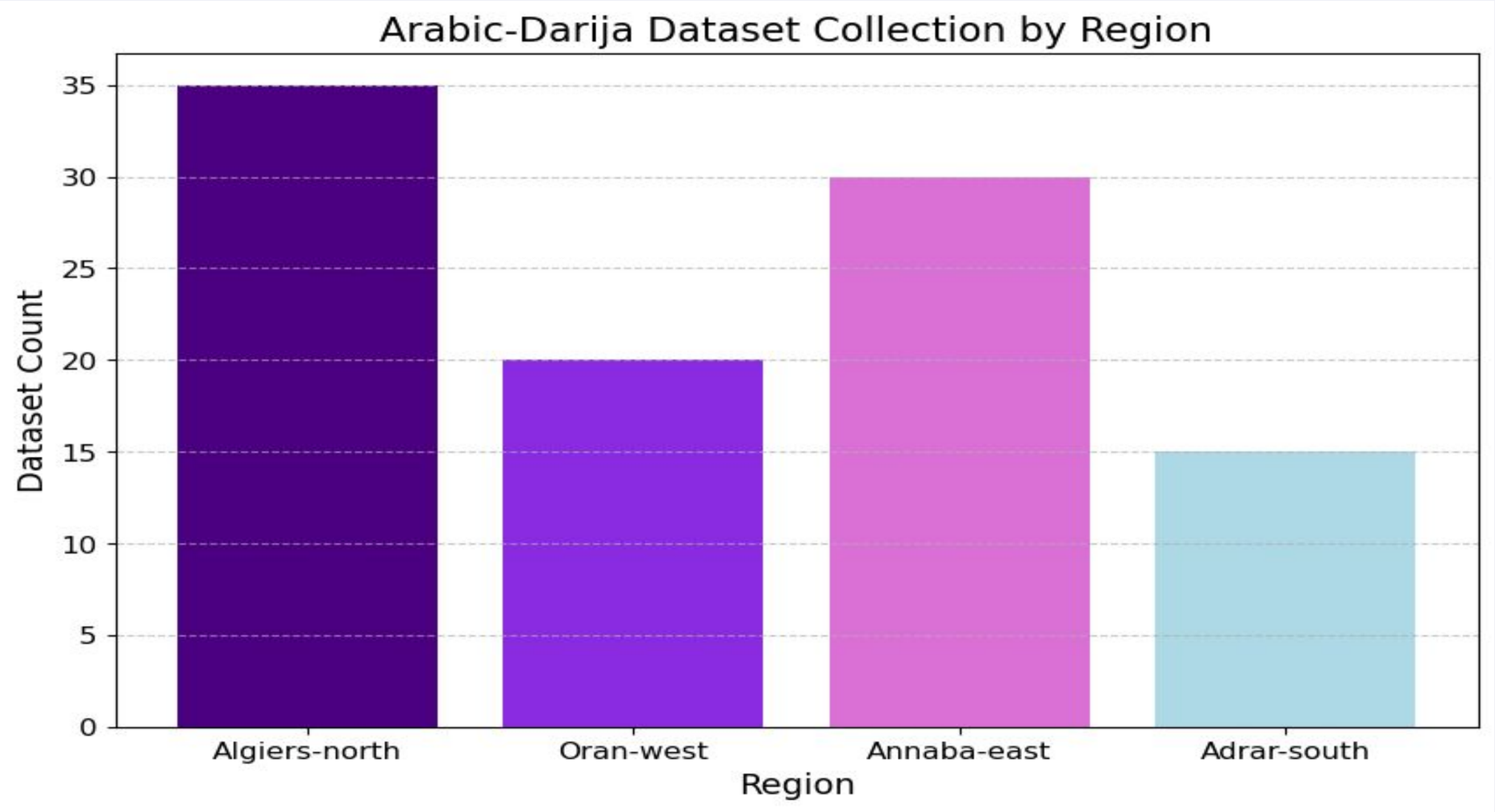


Kabyle data



Darija data

Data Collection - some statistics



Darija data collection by regions

Discussion: Results and Findings

- Training consists of two parts: one for **Kabyle** and one for **Darija**.
- Training was limited to 2 epochs on our local machines, using a maximum of **5%** of the entire dataset, constrained by the need for powerful GPUs.
- During these 2 epochs, the "**Word Error Rate**" (WER) metric showed a high value, around 400, highlighting the need for more epochs and training.
- Testing on the trained model provided by the **fentech** team for Darija showed correct recognition of most words, including those mixed with French and English, with a few errors in word endings or order. The exact WER value for this final trained model was not available, but the tests yielded positive outcomes.
- For the Kabyle dataset, similar behavior was observed, with correct recognition of most text, often displaying results in French letters.

Discussion: Challenges

Data size and format

Data collection presented several challenges, including ensuring the format compatibility with the Whisper model and the requirement for a substantial data size.

Model training requirements

The training process necessitated a powerful GPU, making it impractical to run locally on oour machines

Conclusion: Future Plans

Our work on developing a robust speech-to-text model for Algerian Darija is just the beginning. To further enhance and expand the capabilities of this model, we have outlined several future plans:

Model Enhancement

Continue fine-tuning the Whisper model with additional diverse Darija speech samples to improve accuracy and robustness.

Dataset Expansion

Collect more high-quality recordings from various Algerian regions to cover all dialectal variations and improve model performance.

Multilingual Integration

Incorporate multilingual capabilities to handle code-switching between Darija, Arabic, French, and other languages.

Cultural Exchange

Develop applications that utilize the model to promote North African culture, literature, and media to a global audience

Conclusion: Future Plans

Collaborative Research

Partner with academic and linguistic institutions to further refine the model and explore new applications.

User Feedback:

Implement a feedback mechanism to continually improve the model based on user experiences and suggestions.

Machine Translation

Integrate the speech-to-text model with machine translation systems to facilitate communication and access to North African resources.

Thank you !

Your invaluable contribution have been essential. Thank you for your support and partnership and you are welcome for any **feedback**.