Punjabi Speech to Text and Keyword Searching

1. Brief Description

A. Speech to Text:

The fine-tuned Punjabi Whisper AI (DrishtiSharma/whisper-large-v2-punjabi-700-steps) model takes an audio file as an input and gives the predicted transcript file (utf-8) as an output.

B. Keyword and Semantic Search (GUI):

This is a basic GUI which takes a Text File as input, and provides options for keyword search (exact or approx) and semantic search (same meaning words/sentences implying same meaning) for checking if the file is suspicious. Exact matches are highlighted green, approx matches are highlighted yellow.

2. System Requirements

A. Speech to Text:

- OS: Linux
  In the case of Windows, use remote vs-code and choose the system as Linux.
  In the case of MacOS, use the Mac terminal.
- Programming Language: Python 3.8
- Libraries: PyTorch, transformers (Hugging Face), pandas, datasets

B. Keyword and Semantic Search:

It is a python code which can run on any system.
- OS: Windows/Linux/macOS
- Programming Language: Python 3.10
- Libraries: tkinter, fuzzywuzzy, os, re, sentence_transformers. torch
- ML models: paraphrase-multilingual-MiniLM-L12-v2 (hugging face, present in sentence_transformers)

3. Installation Instructions

A. Speech to Text:

➢ Step-by-step guide for getting access to GPU container:

- Open the link: Submit Ticket. Click on "NVIDIA GPU Servers".
- Fill the details, change the priority to "High" and in the subject and message sections, mention your supervisor's name.
- The IT section will respond within 2-3 working days.
- Tip: Get the GPU in your supervisor's name to get more computational resources.

➢ Step-by-step guide to setup remote VS code (for Windows):

- Follow the link: Remote VS Code setup
- For Mac, simply use the Mac terminal with vim editor.

➢ Step-by-step guide to setup VPN (in case you are not using IITRPR wifi):

- Create an LDAP account.

- Follow the exact steps given in the link: VPN setup

➢ Step-by-step guide to install all necessary dependencies for remote system:

- Transferring files/folders from local system to remote system:
  scp file_path userame@ip:~
- Setup virtual environment:
  sudo apt install python3-venv
  python3 -m venv venv_name
  source venv_name/bin/activate
- Installing libraries:
  - Create a requirements.txt file containing names of all the libraries required.
  pip install -r requirements.txt
- Check real-time GPU usage:
  watch -n 1 nvidia-smi
  Note: Don't forget to press Esc or kill the local terminal, otherwise it will keep running.

B. Keyword and Semantic Search:

Make sure you have the latest version of python (code is in python 3.10) with all the libraries mentioned earlier installed. You can run the GUI python on a local machine (eg:vscode), it opens a local window for GUI.

4. How to Run the Software

A. Speech to Text:

Step-by-step instructions to run files on the remote system:

- Tmux terminal (so that code keeps running in the background):
  - To launch a new tmux terminal—-> tmux
  - Activate your virtual environment—--> source venv_name/bin/activate
  - Note the tmux terminal number (say for example 2), shown in square brackets at the left end of the green line at the bottom of the terminal.
  - Run the desired file.
  - Now when you get back to your laptop after a long time, the tmux terminal would have disappeared. Don't panic.
  - Reopen tmux terminal running in the background—--> tmux attach-session -t 2 (Here 2 is the terminal number)
  - To scroll through the tmux terminal (Windows)—-> Press Ctrl+b followed by [
  - To scroll through the tmux terminal (Mac)—-> Press Control+b
  - To exit scrolling mode in tmux terminal—---> Press Esc
  - Never ever type 'exit' in a tmux terminal; it closes the tmux terminal permanently

- For training on LDCIL dataset:
  cd BTP/whisper_ash
  python3 final.py

- For testing on LDCIL dataset:
  cd BTP/whisper_ash
  python3 testing.py

- For testing original Drishti Sharma model on police files:
  cd BTP/whisper_ash/police_files
  python3 original_police.py

- For testing further fine-tuned Drishti Sharma model on police files:
  - cd BTP/whisper_ash/police_files
  - python3 police.py

Switching configurations:

- testing.py (default is fine-tuned):

  ➢ To test the original Drishti Sharma model on the LDCIL test dataset,

  - Comment out the following lines:
  Line 12: model_path = './my_finetuned_model' #--> to test fine-tuned
  Line 25: transcript_output_dir = os.path.join(dataset_path, 'predicted_transcripts_finetune_ldcil') #--> for saving fine tuned predicted transcripts
  Line 37: processor = WhisperProcessor.from_pretrained(model_path) #--> to test fine tuned
  Line 39: model = WhisperForConditionalGeneration.from_pretrained(model_path).to(device) #--> to test fine tuned

  - Uncomment the following lines:
  Line 13: # model_name = "DrishtiSharma/whisper-large-v2-punjabi-700-steps" #--> to test original
  Line 26: #transcript_output_dir = os.path.join(dataset_path, 'predicted_transcripts_original_ldcil') #--> for saving original predicted transcripts
  Line 38: #processor = WhisperProcessor.from_pretrained(model_name) #--> to test original
  Line 40: #model = WhisperForConditionalGeneration.from_pretrained(model_name).to(device) #--> to test original

- original_police.py (default is for High Quality files):

  ➢ To test the model (original or fine-tuned) on Medium or Low quality files,

  - Replace ‘High’ by ‘Medium’ or ‘Low’ in the following lines:
  Line 46: police_files_path=os.path.join(BASE_DIR,"whisper_ash/police_files/High/")
  Line 47: output_folder = os.path.join(BASE_DIR, 'whisper_ash/police_files/predicted_transcripts_original/HIGH_sentences')
  Line 248: txt_file_name = f"High_{district}_Punjabi_{audio_file_name}.txt"

- police.py (default is for High Quality files):

  ➢ To test the model (original or fine-tuned) on Medium or Low quality files,

  - Replace ‘High’ by ‘Medium’ or ‘Low’ in the following lines:
  Line 53: police_files_path=os.path.join(BASE_DIR,"whisper_ash/police_files/High/")
  Line 54: output_folder = os.path.join(BASE_DIR, 'whisper_ash/police_files/predicted_transcripts_finetuned/HIGH_sentences')
  Line 305: txt_file_name = f"High_{district}_Punjabi_{audio_file_name}.txt"

Output files:

- Predicted transcripts on LDCIL test dataset:
  - Original Drishti Sharma model:
  cd BTP/whisper_ash/predicted_transcripts_original_ldcil
  - Fine-tuned model:
  cd BTP/whisper_ash/predicted_transcripts_finetune_ldcil

- Predicted transcripts on police files:

□ - Original Drishti Sharma model:
□ cd BTP/whisper_ash/police_files/predicted_transcripts_original/HIGH_sentences
□ (Similarly for MEDIUM_sentences and LOW_sentences)
□ - Fine-tuned model:
□ cd BTP/whisper_ash/police_files/predicted_transcripts_finetuned/HIGH_sentences
□ (Similarly for MEDIUM_sentences and LOW_sentences)
□

B. Keyword and Semantic Search:

Step-by-step instructions to run the GUI:

Run the python code on your local machine (we used VScode). A new window (GUI) opens in 8-10 seconds (as models need to be called).

- First page has a “Select File” option. Select the Punjabi text file from the local machine.
- As soon as you select the file, the second page for Exact and Fuzzy (approx) keyword search opens. It has:
  - “Search” bar and button, where you can enter any number of words separated by comma.
  - Boxes displaying number of exact and fuzzy matches for given keywords (new box for each keyword).
  - Two text files Original and Highlighted (where exact and fuzzy matches are highlighted green and yellow respectively) with scroll bars.
  - “Save Output” button to save the highlighted text file (in the same folder as the original file, with name originalFileName_text.txt). Here exact matches are underlined, and fuzzy matches are in bold.
  - “Back” button to go to the first page and select a new file.
  - “Semantic Search” button to go to the next page for semantic search on the same selected text file.
- As soon as you select “Semantic Search”, the third page for Semantic Search opens. It has:
  - “Search” bar where you can enter any number of words separated by comma.
  - “Run Semantic Search” button to run semantic analysis of entered keywords on the text file selected in the first page. It might take time depending on the number of keywords entered and the length of the text file.
  - Text Box with scroll bar to print how much the text file implies meaning related to the entered keywords. It gives scores for each keyword (>0.3 is a good match) along with a number of exact and fuzzy matches in brackets.
  - “Back” button to go to the second page.

Note:

➔ Default keyword for Second page is “ਵੋਟ”. This can be changed.

➔ Default keywords for Third page are:
“ਪਿਸਤੌਲ, ਸਪਲਾਇਰ, ਡ੍ਰੌਪ-ਆਫ਼, ਹਥਿਆਰ, ਦਵਾਈ, ਰਾਈਫਲ, ਡ੍ਰੌਪ, ਕੈਦੀ”
These can also be changed.

5. Directory Structure

A. Speech to Text:

An overview of the directory structure in remote system:

```
home/
|
bhumika/
|
BTP/
|
whisper_ash/
|___concatenated_clips/
|       |__Audio_File_1.wav
|       |__Audio_File_2.wav
|       |       :
|       |       :
|       |__Audio_File_2792.wav
|
|___my_finetuned_model/
|
|___police_files/
|      |__High/
|      |     |__Amritsar/
|      |     |     |__Hindi/
|      |     |     |__Punjabi/
|      |     |         |__High Quality 1.WAV
|      |     |         |__HIGH QUALAITY.WAV
|      |     |__Kapurthala/
|      |     |     |__Hindi/
|      |     |     |__Punjabi/
|      |     |         |__high 1.wav
|      |     |         |__high.wav
|      |     |__Patiala/
|      |     |     |__Hindi/
|      |     |     |__Punjabi/
|      |     |         |__high.WAV
|      |     |         |__high1.WAV
|      |     |__Ropar/
|      |     |     |__Hindi/
|      |     |     |__Punjabi/
|      |     |         |__high.WAV
|      |     |         |__high1.WAV
|      |     |
|      |__Low/
|      |     |__Amritsar/
|      |     |     |__Hindi/
|      |     |     |__Punjabi/
|      |     |         |__LOW QUALITY (2).WAV
|      |     |         |__LOW QUALITY.WAV
|      |     |__Kapurthala/
|      |     |     |__Hindi/
|      |     |     |__Punjabi/
|      |     |         |__low.wav
|      |     |         |__low1.wav
|      |     |__Patiala/
|      |     |     |__Hindi/
|      |     |     |__Punjabi/
|      |     |         |__low.WAV
|     |    |       |__low1.WAV
|     |    |__Ropar/
|     |    |    |__Hindi/
|     |    |    |__Punjabi/
|     |    |       |__low.WAV
|     |    |       |__low1.WAV
|     |    |
|     |__Medium/
|     |    |__Amritsar/
|     |    |    |__Hindi/
|     |    |    |__Punjabi/
|     |    |       |__medium quality 1.WAV
|     |    |       |__Medium.WAV
|     |    |__Kapurthala/
|     |    |    |__Hindi/
|     |    |    |__Punjabi/
|     |    |       |__med.wav
|     |    |       |__med1.wav
|     |    |__Patiala/
|     |    |    |__Hindi/
|     |    |    |__Punjabi/
|     |    |       |__med.WAV
|     |    |       |__med1.WAV
|     |    |__Ropar/
|     |    |    |__Hindi/
|     |    |    |__Punjabi/
|     |    |       |__med.WAV
|     |    |       |__med1.WAV
|     |    |
|     |__predicted_transcripts_finetuned/
|     |    |__HIGH_sentences/
|     |    |    |__High_Amritsar_Punjabi_High Quality 1.txt
|     |    |    |__High_Amritsar_Punjabi_HIGH QULAITY.txt
|     |    |    |__High_Kapurthala_Punjabi_high 1.txt
|     |    |    |__High_Kapurthala_Punjabi_high.txt
|     |    |    |__High_Patiala_Punjabi_high.txt
|     |    |    |__High_Patiala_Punjabi_high1.txt
|     |    |    |__High_Ropar_Punjabi_high.txt
|     |    |    |__High_Ropar_Punjabi_high1.txt
|     |    |
|     |    |__LOW_sentences/
|     |    |    |__Low_Amritsar_Punjabi_LOW QUALITY (2).txt
|     |    |    |__Low_Amritsar_Punjabi_LOW QUALITY.txt
|     |    |    |__Low_Kapurthala_Punjabi_low.txt
|     |    |    |__Low_Kapurthala_Punjabi_low1.txt
|     |    |    |__Low_Patiala_Punjabi_low.txt
|     |    |    |__Low_Patiala_Punjabi_low1.txt
|     |    |    |__Low_Ropar_Punjabi_low.txt
|     |    |    |__Low_Ropar_Punjabi_low1.txt
|     |    |
|     |    |__MEDIUM_sentences/
|     |    |    |__Medium_Amritsar_Punjabi_medium quality 1.txt
|     |    |    |__Medium_Amritsar_Punjabi_Medium.txt
|     |    |    |__Medium_Kapurthala_Punjabi_med.txt
|     |    |    |__Medium_Kapurthala_Punjabi_med1.txt
|     |    |    |__Medium_Patiala_Punjabi_med.txt
|     |    |    |__Medium_Patiala_Punjabi_med1.txt
|     |    |    |__ Medium_Ropar_Punjabi_med.txt
|     |    |    |__ Medium_Ropar_Punjabi_med1.txt
|     |     |
|     |__predicted_transcripts_original/
|     |     |__HIGH_sentences/
|     |     |     |__High_Amritsar_Punjabi_High Quality 1.txt
|     |     |     |__High_Amritsar_Punjabi_HIGH QULAITY.txt
|     |     |     |__High_Kapurthala_Punjabi_high 1.txt
|     |     |     |__High_Kapurthala_Punjabi_high.txt
|     |     |     |__High_Patiala_Punjabi_high.txt
|     |     |     |__High_Patiala_Punjabi_high1.txt
|     |     |     |__High_Ropar_Punjabi_high.txt
|     |     |     |__High_Ropar_Punjabi_high1.txt
|     |     |
|     |     |__LOW_sentences/
|     |     |     |__Low_Amritsar_Punjabi_LOW QUALITY (2).txt
|     |     |     |__Low_Amritsar_Punjabi_LOW QUALITY.txt
|     |     |     |__Low_Kapurthala_Punjabi_low.txt
|     |     |     |__Low_Kapurthala_Punjabi_low1.txt
|     |     |     |__Low_Patiala_Punjabi_low.txt
|     |     |     |__Low_Patiala_Punjabi_low1.txt
|     |     |     |__Low_Ropar_Punjabi_low.txt
|     |     |     |__Low_Ropar_Punjabi_low1.txt
|     |     |
|     |     |__MEDIUM_sentences/
|     |     |     |__Medium_Amritsar_Punjabi_medium quality 1.txt
|     |     |     |__Medium_Amritsar_Punjabi_Medium.txt
|     |     |     |__Medium_Kapurthala_Punjabi_med.txt
|     |     |     |__Medium_Kapurthala_Punjabi_med1.txt
|     |     |     |__Medium_Patiala_Punjabi_med.txt
|     |     |     |__Medium_Patiala_Punjabi_med1.txt
|     |     |     |__ Medium_Ropar_Punjabi_med.txt
|     |     |     |__ Medium_Ropar_Punjabi_med1.txt
|     |     |
|     |__original_police.py
|     |__police.py
|___|
|___predicted_transcripts_finetune_ldcil/
|___predicted_transcripts_original_ldcil/
|___final.py
|___testing.py
|___train.txt
|___text.txt
|
requirements.txt
|
btp_venv/
```

B. Keyword and Semantic Search:

You can run the python code anywhere on your local system. Output text file (highlighted) is saved in the same directory as the input text file if the “Save” option is clicked.

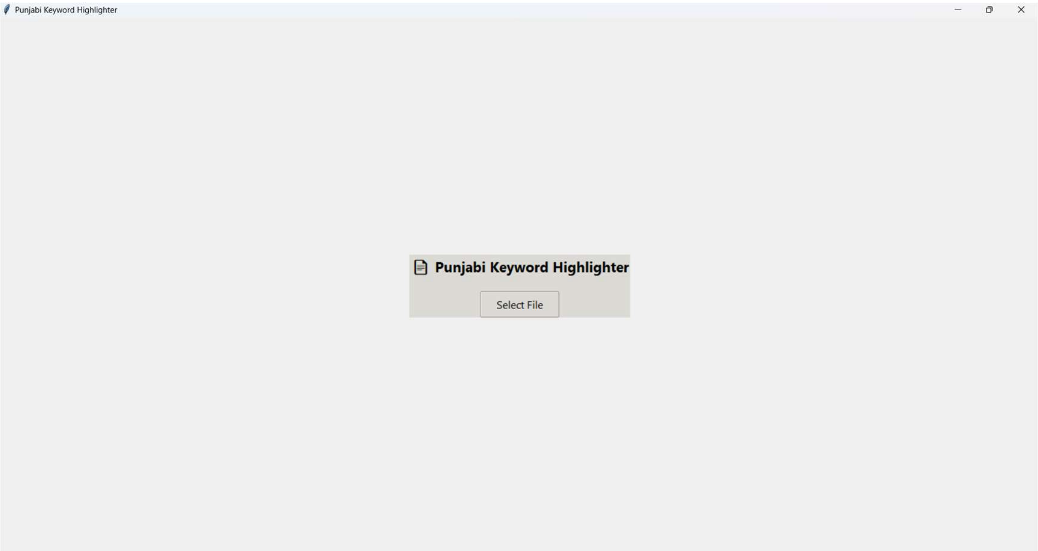
6. Sample Inputs and Outputs

A. Speech to Text:

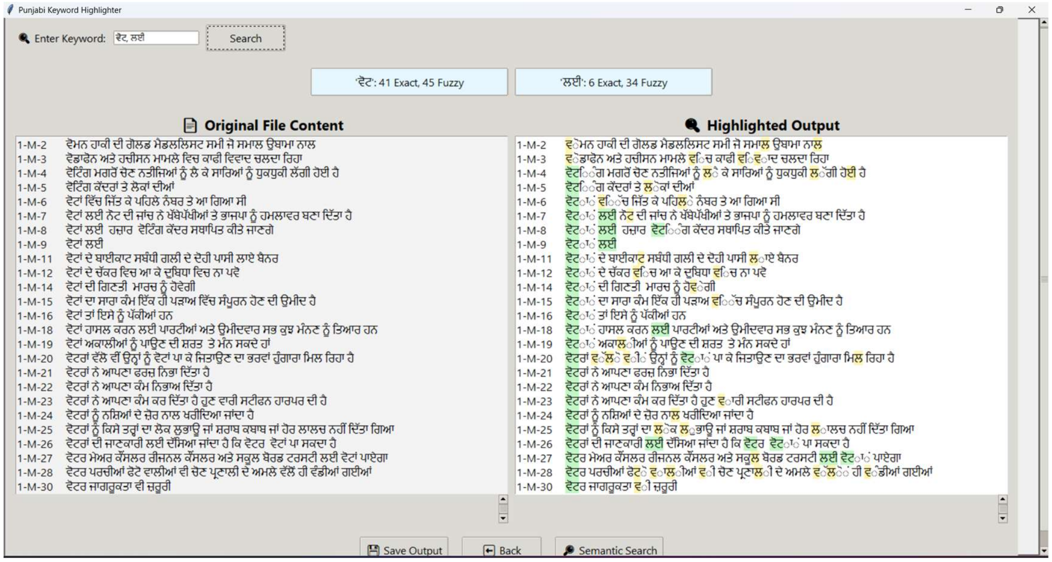
- Sample Input: Any audio file (.wav or .WAV)
- Sample Output: A .txt file containing the Punjabi transcript

B. Keyword and Semantic Search:

Following are step by step screenshots of the GUI for reference:



File selected: M1-M50.txt



Enter Keyword: ਵੋਟ, ਲਈ

'ਵੋਟ': 41 Exact, 45 Fuzzy

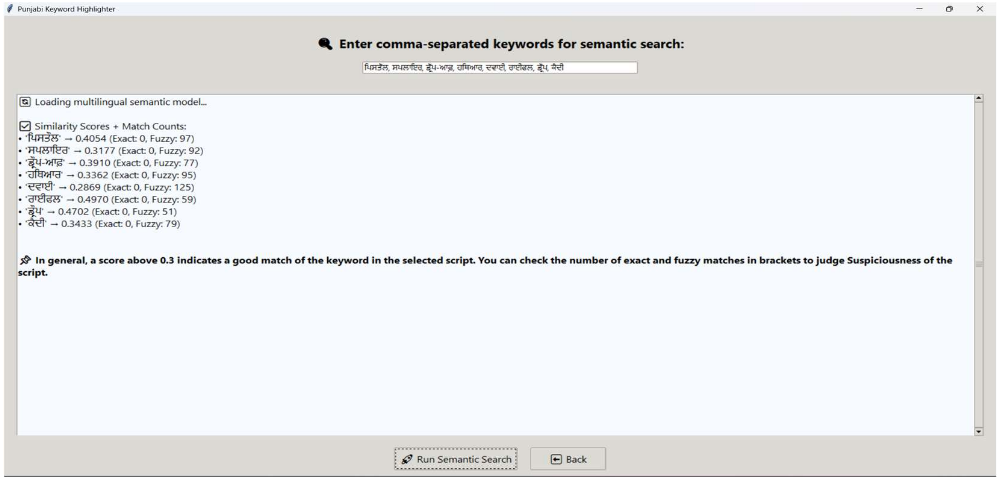
'ਲਈ': 6 Exact, 34 Fuzzy

Original File Content

```
1-M-2   ਵੋਮਨ ਹਾਕੀ ਦੀ ਗੋਲਡ ਮੈਡਲਲਿਸਟ ਸਮੀ ਜੋ ਸਮਾਲ ਉਥਾਮਾ ਨਾਲ
1-M-3   ਵੋਡਾਫੋਨ ਅਤੇ ਹਚੀਸਨ ਮਾਮਲੇ ਵਿਚ ਕਾਫੀ ਵਿਵਾਦ ਚਲਦਾ ਰਿਹਾ
1-M-4   ਵੋਟਿੰਗ ਮਗਰੋਂ ਚੋਣ ਨਤੀਜਿਆਂ ਨੂੰ ਲੈ ਕੇ ਸਾਰਿਆਂ ਨੂੰ ਧੁਕਧੁਕੀ ਲੱਗੀ ਹੋਈ ਹੈ
1-M-5   ਵੋਟਿੰਗ ਕੇਂਦਰਾਂ ਤੇ ਲੋਕਾਂ ਦੀਆਂ
1-M-6   ਵੋਟਾਂ ਵਿੱਚ ਜਿੱਤ ਕੇ ਪਹਿਲੇ ਨੰਬਰ ਤੇ ਆ ਗਿਆ ਸੀ
1-M-7   ਵੋਟਾਂ ਲਈ ਨੋਟ ਦੀ ਜਾਂਚ ਨੇ ਖੱਬੇਪੱਖੀਆਂ ਤੇ ਭਾਜਪਾ ਨੂੰ ਹਮਲਾਵਰ ਬਣਾ ਦਿੱਤਾ ਹੈ
1-M-8   ਵੋਟਾਂ ਲਈ  ਹਜ਼ਾਰ  ਵੋਟਿੰਗ ਕੇਂਦਰ ਸਥਾਪਿਤ ਕੀਤੇ ਜਾਣਗੇ
1-M-9   ਵੋਟਾਂ ਲਈ
1-M-11  ਵੋਟਾਂ ਦੇ ਬਾਈਕਾਟ ਸਬੰਧੀ ਗਲੀ ਦੇ ਦੋਹੀ ਪਾਸੀ ਲਾਏ ਬੈਨਰ
1-M-12  ਵੋਟਾਂ ਦੇ ਚੱਕਰ ਵਿਚ ਆ ਕੇ ਦੁਬਿਧਾ ਵਿਚ ਨਾ ਪਵੋ
1-M-14  ਵੋਟਾਂ ਦੀ ਗਿਣਤੀ  ਮਾਰਚ ਨੂੰ ਹੋਵੇਗੀ
1-M-15  ਵੋਟਾਂ ਦਾ ਸਾਰਾ ਕੰਮ ਇੱਕ ਹੀ ਪੜਾਅ ਵਿੱਚ ਸੰਪੂਰਨ ਹੋਣ ਦੀ ਉਮੀਦ ਹੈ
1-M-16  ਵੋਟਾਂ ਤਾਂ ਇਸੇ ਨੂੰ ਪੱਕੀਆਂ ਹਨ
1-M-18  ਵੋਟਾਂ ਹਾਸਲ ਕਰਨ ਲਈ ਪਾਰਟੀਆਂ ਅਤੇ ਉਮੀਦਵਾਰ ਸਭ ਕੁਝ ਮੰਨਣ ਨੂੰ ਤਿਆਰ ਹਨ
1-M-19  ਵੋਟਾਂ ਅਕਾਲੀਆਂ ਨੂੰ ਪਾਉਣ ਦੀ ਸ਼ਰਤ  ਤੇ ਮੰਨ ਸਕਦੇ ਹਾਂ
1-M-20  ਵੋਟਰਾਂ ਵੱਲੋ ਵੀਂ ਉਨ੍ਹਾਂ ਨੂੰ ਵੋਟਾਂ ਪਾ ਕੇ ਜਿਤਾਉਣ ਦਾ ਭਰਵਾਂ ਹੁੰਗਾਰਾ ਮਿਲ ਰਿਹਾ ਹੈ
1-M-21  ਵੋਟਰਾਂ ਨੇ ਆਪਣਾ ਫਰਜ਼ ਨਿਭਾ ਦਿੱਤਾ ਹੈ
1-M-22  ਵੋਟਰਾਂ ਨੇ ਆਪਣਾ ਕੰਮ ਨਿਭਾਅ ਦਿੱਤਾ ਹੈ
1-M-23  ਵੋਟਰਾਂ ਨੇ ਆਪਣਾ ਕੰਮ ਕਰ ਦਿੱਤਾ ਹੈ ਹੁਣ ਵਾਰੀ ਸਟੀਫਨ ਹਾਰਪਰ ਦੀ ਹੈ
1-M-24  ਵੋਟਰਾਂ ਨੂੰ ਨਸ਼ਿਆਂ ਦੇ ਜ਼ੋਰ ਨਾਲ ਖਰੀਦਿਆ ਜਾਂਦਾ ਹੈ
1-M-25  ਵੋਟਰਾਂ ਨੂੰ ਕਿਸੇ ਤਰ੍ਹਾਂ ਦਾ ਲੋਕ ਲੁਭਾਊ ਜਾਂ ਸ਼ਰਾਬ ਕਬਾਬ ਜਾਂ ਹੋਰ ਲਾਲਚ ਨਹੀਂ ਦਿੱਤਾ ਗਿਆ
1-M-26  ਵੋਟਰਾਂ ਦੀ ਜਾਣਕਾਰੀ ਲਈ ਦੱਸਿਆ ਜਾਂਦਾ ਹੈ ਕਿ ਵੋਟਰ  ਵੋਟਾਂ ਪਾ ਸਕਦਾ ਹੈ
1-M-27  ਵੋਟਰ ਮੇਅਰ ਕੌਂਸਲਰ ਰੀਜਨਲ ਕੌਂਸਲਰ ਅਤੇ ਸਕੂਲ ਬੋਰਡ ਟਰਸਟੀ ਲਈ ਵੋਟਾਂ ਪਾਏਗਾ
1-M-28  ਵੋਟਰ ਪਰਚੀਆਂ ਫੋਟੋ ਵਾਲੀਆਂ ਵੀ ਚੋਣ ਪ੍ਰਣਾਲੀ ਦੇ ਅਮਲੇ ਵੱਲੋਂ ਹੀ ਵੰਡੀਆਂ ਗਈਆਂ
1-M-30  ਵੋਟਰ ਜਾਗਰੂਕਤਾ ਵੀ ਜ਼ਰੂਰੀ
```

Highlighted Output

```
1-M-2   ਵੋਮਨ ਹਾਕੀ ਦੀ ਗੋਲਡ ਮੈਡਲਲਿਸਟ ਸਮੀ ਜੋ ਸਮਾਲ ਉਥਾਮਾ ਨਾਲ
1-M-3   ਵੋਡਾਫੋਨ ਅਤੇ ਹਚੀਸਨ ਮਾਮਲੇ ਵਿਚ ਕਾਫੀ ਵਿਵਾਦ ਚਲਦਾ ਰਿਹਾ
1-M-4   ਵੋਟਿੰਗ ਮਗਰੋਂ ਚੋਣ ਨਤੀਜਿਆਂ ਨੂੰ ਲੈ ਕੇ ਸਾਰਿਆਂ ਨੂੰ ਧੁਕਧੁਕੀ ਲੱਗੀ ਹੋਈ ਹੈ
1-M-5   ਵੋਟਿੰਗ ਕੇਂਦਰਾਂ ਤੇ ਲੋਕਾਂ ਦੀਆਂ
1-M-6   ਵੋਟਾਂ ਵਿੱਚ ਜਿੱਤ ਕੇ ਪਹਿਲੇ ਨੰਬਰ ਤੇ ਆ ਗਿਆ ਸੀ
1-M-7   ਵੋਟਾਂ ਲਈ ਨੋਟ ਦੀ ਜਾਂਚ ਨੇ ਖੱਬੇਪੱਖੀਆਂ ਤੇ ਭਾਜਪਾ ਨੂੰ ਹਮਲਾਵਰ ਬਣਾ ਦਿੱਤਾ ਹੈ
1-M-8   ਵੋਟਾਂ ਲਈ  ਹਜ਼ਾਰ  ਵੋਟਿੰਗ ਕੇਂਦਰ ਸਥਾਪਿਤ ਕੀਤੇ ਜਾਣਗੇ
1-M-9   ਵੋਟਾਂ ਲਈ
1-M-11  ਵੋਟਾਂ ਦੇ ਬਾਈਕਾਟ ਸਬੰਧੀ ਗਲੀ ਦੇ ਦੋਹੀ ਪਾਸੀ ਲਾਏ ਬੈਨਰ
1-M-12  ਵੋਟਾਂ ਦੇ ਚੱਕਰ ਵਿਚ ਆ ਕੇ ਦੁਬਿਧਾ ਵਿਚ ਨਾ ਪਵੋ
1-M-14  ਵੋਟਾਂ ਦੀ ਗਿਣਤੀ  ਮਾਰਚ ਨੂੰ ਹੋਵੇਗੀ
1-M-15  ਵੋਟਾਂ ਦਾ ਸਾਰਾ ਕੰਮ ਇੱਕ ਹੀ ਪੜਾਅ ਵਿੱਚ ਸੰਪੂਰਨ ਹੋਣ ਦੀ ਉਮੀਦ ਹੈ
1-M-16  ਵੋਟਾਂ ਤਾਂ ਇਸੇ ਨੂੰ ਪੱਕੀਆਂ ਹਨ
1-M-18  ਵੋਟਾਂ ਹਾਸਲ ਕਰਨ ਲਈ ਪਾਰਟੀਆਂ ਅਤੇ ਉਮੀਦਵਾਰ ਸਭ ਕੁਝ ਮੰਨਣ ਨੂੰ ਤਿਆਰ ਹਨ
1-M-19  ਵੋਟਾਂ ਅਕਾਲੀਆਂ ਨੂੰ ਪਾਉਣ ਦੀ ਸ਼ਰਤ  ਤੇ ਮੰਨ ਸਕਦੇ ਹਾਂ
1-M-20  ਵੋਟਰਾਂ ਵੱਲੋ ਵੀਂ ਉਨ੍ਹਾਂ ਨੂੰ ਵੋਟਾਂ ਪਾ ਕੇ ਜਿਤਾਉਣ ਦਾ ਭਰਵਾਂ ਹੁੰਗਾਰਾ ਮਿਲ ਰਿਹਾ ਹੈ
1-M-21  ਵੋਟਰਾਂ ਨੇ ਆਪਣਾ ਫਰਜ਼ ਨਿਭਾ ਦਿੱਤਾ ਹੈ
1-M-22  ਵੋਟਰਾਂ ਨੇ ਆਪਣਾ ਕੰਮ ਨਿਭਾਅ ਦਿੱਤਾ ਹੈ
1-M-23  ਵੋਟਰਾਂ ਨੇ ਆਪਣਾ ਕੰਮ ਕਰ ਦਿੱਤਾ ਹੈ ਹੁਣ ਵਾਰੀ ਸਟੀਫਨ ਹਾਰਪਰ ਦੀ ਹੈ
1-M-24  ਵੋਟਰਾਂ ਨੂੰ ਨਸ਼ਿਆਂ ਦੇ ਜ਼ੋਰ ਨਾਲ ਖਰੀਦਿਆ ਜਾਂਦਾ ਹੈ
1-M-25  ਵੋਟਰਾਂ ਨੂੰ ਕਿਸੇ ਤਰ੍ਹਾਂ ਦਾ ਲੋਕ ਲੁਭਾਊ ਜਾਂ ਸ਼ਰਾਬ ਕਬਾਬ ਜਾਂ ਹੋਰ ਲਾਲਚ ਨਹੀਂ ਦਿੱਤਾ ਗਿਆ
1-M-26  ਵੋਟਰਾਂ ਦੀ ਜਾਣਕਾਰੀ ਲਈ ਦੱਸਿਆ ਜਾਂਦਾ ਹੈ ਕਿ ਵੋਟਰ  ਵੋਟਾਂ ਪਾ ਸਕਦਾ ਹੈ
1-M-27  ਵੋਟਰ ਮੇਅਰ ਕੌਂਸਲਰ ਰੀਜਨਲ ਕੌਂਸਲਰ ਅਤੇ ਸਕੂਲ ਬੋਰਡ ਟਰਸਟੀ ਲਈ ਵੋਟਾਂ ਪਾਏਗਾ
1-M-28  ਵੋਟਰ ਪਰਚੀਆਂ ਫੋਟੋ ਵਾਲੀਆਂ ਵੀ ਚੋਣ ਪ੍ਰਣਾਲੀ ਦੇ ਅਮਲੇ ਵੱਲੋਂ ਹੀ ਵੰਡੀਆਂ ਗਈਆਂ
1-M-30  ਵੋਟਰ ਜਾਗਰੂਕਤਾ ਵੀ ਜ਼ਰੂਰੀ
```



Enter comma-separated keywords for semantic search:

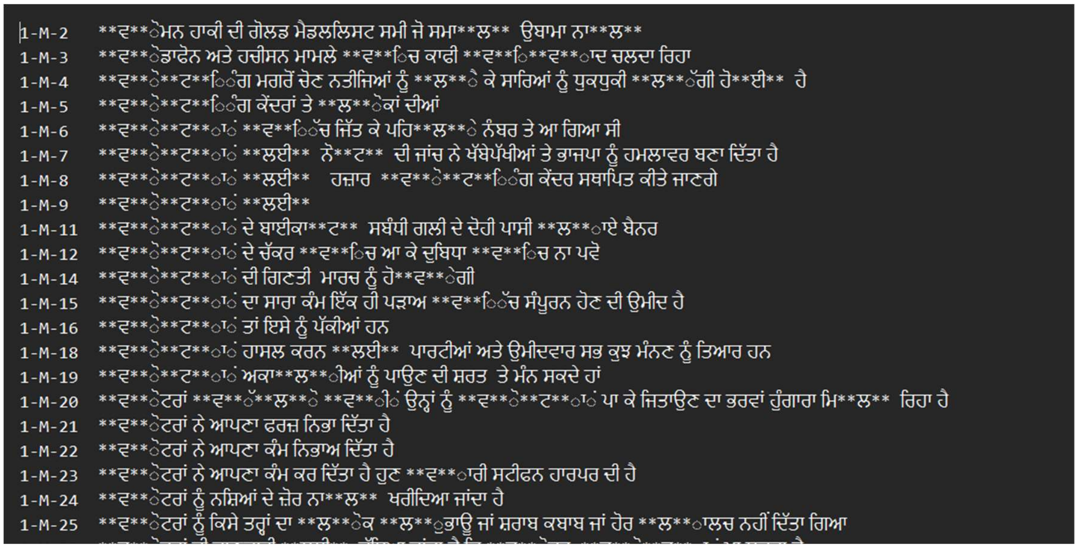
ਪਿਸਤੌਲ, ਸਪਲਾਇਰ, ਡ੍ਰੌਪ-ਆਫ਼, ਹਥਿਆਰ, ਦਵਾਈ, ਰਾਈਫਲ, ਡ੍ਰੌਪ, ਕੈਦੀ

```
Loading multilingual semantic model...

☑ Similarity Scores + Match Counts:
• 'ਪਿਸਤੌਲ' → 0.4054 (Exact: 0, Fuzzy: 97)
• 'ਸਪਲਾਇਰ' → 0.3177 (Exact: 0, Fuzzy: 92)
• 'ਡ੍ਰੌਪ-ਆਫ਼' → 0.3910 (Exact: 0, Fuzzy: 77)
• 'ਹਥਿਆਰ' → 0.3362 (Exact: 0, Fuzzy: 95)
• 'ਦਵਾਈ' → 0.2869 (Exact: 0, Fuzzy: 125)
• 'ਰਾਈਫਲ' → 0.4970 (Exact: 0, Fuzzy: 59)
• 'ਡ੍ਰੌਪ' → 0.4702 (Exact: 0, Fuzzy: 51)
• 'ਕੈਦੀ' → 0.3433 (Exact: 0, Fuzzy: 79)
```

In general, a score above 0.3 indicates a good match of the keyword in the selected script. You can check the number of exact and fuzzy matches in brackets to judge Suspiciousness of the script.

Text File saved (M1-M50_text.txt):



```
1-M-2    **ਵ**ੋਮਨ ਹਾਕੀ ਦੀ ਗੋਲਡ ਮੈਡਲਲਿਸਟ ਸਮੀ ਜੋ ਸਮਾ**ਲ**  ਉਬਾਮਾ ਨਾ**ਲ**
1-M-3    **ਵ**ੋਡਾਫੋਨ ਅਤੇ ਹਚੀਸਨ ਮਾਮਲੇ **ਵ**ਿਚ ਕਾਫੀ **ਵ**ਿ**ਵ**ਾਦ ਚਲਦਾ ਰਿਹਾ
1-M-4    **ਵ**ੋ**ਟ**ਿੰਗ ਮਗਰੋਂ ਚੋਣ ਨਤੀਜਿਆਂ ਨੂੰ **ਲ**ੈ ਕੇ ਸਾਰਿਆਂ ਨੂੰ ਧੁਕਧੁਕੀ **ਲ**ੱਗੀ ਹੋ**ਈ**  ਹੈ
1-M-5    **ਵ**ੋ**ਟ**ਿੰਗ ਕੇਂਦਰਾਂ ਤੇ **ਲ**ੋਕਾਂ ਦੀਆਂ
1-M-6    **ਵ**ੋ**ਟ**ਾਂ **ਵ**ਿੱਚ ਜਿੱਤ ਕੇ ਪਹਿ**ਲ**ੇ ਨੰਬਰ ਤੇ ਆ ਗਿਆ ਸੀ
1-M-7    **ਵ**ੋ**ਟ**ਾਂ **ਲਈ**  ਨੋ**ਟ**  ਦੀ ਜਾਂਚ ਨੇ ਖੱਬੇਪੱਖੀਆਂ ਤੇ ਭਾਜਪਾ ਨੂੰ ਹਮਲਾਵਰ ਬਣਾ ਦਿੱਤਾ ਹੈ
1-M-8    **ਵ**ੋ**ਟ**ਾਂ **ਲਈ**    ਹਜ਼ਾਰ  **ਵ**ੋ**ਟ**ਿੰਗ ਕੇਂਦਰ ਸਥਾਪਿਤ ਕੀਤੇ ਜਾਣਗੇ
1-M-9    **ਵ**ੋ**ਟ**ਾਂ **ਲਈ**
1-M-11   **ਵ**ੋ**ਟ**ਾਂ ਦੇ ਬਾਈਕਾ**ਟ**  ਸਬੰਧੀ ਗਲੀ ਦੇ ਦੋਹੀ ਪਾਸੀ **ਲ**ਾਏ ਬੈਨਰ
1-M-12   **ਵ**ੋ**ਟ**ਾਂ ਦੇ ਚੱਕਰ **ਵ**ਿਚ ਆ ਕੇ ਦੁਬਿਧਾ **ਵ**ਿਚ ਨਾ ਪਵੋ
1-M-14   **ਵ**ੋ**ਟ**ਾਂ ਦੀ ਗਿਣਤੀ  ਮਾਰਚ ਨੂੰ ਹੋ**ਵ**ੇਗੀ
1-M-15   **ਵ**ੋ**ਟ**ਾਂ ਦਾ ਸਾਰਾ ਕੰਮ ਇੱਕ ਹੀ ਪੜਾਅ **ਵ**ਿੱਚ ਸੰਪੂਰਨ ਹੋਣ ਦੀ ਉਮੀਦ ਹੈ
1-M-16   **ਵ**ੋ**ਟ**ਾਂ ਤਾਂ ਇਸੇ ਨੂੰ ਪੱਕੀਆਂ ਹਨ
1-M-18   **ਵ**ੋ**ਟ**ਾਂ ਹਾਸਲ ਕਰਨ **ਲਈ**  ਪਾਰਟੀਆਂ ਅਤੇ ਉਮੀਦਵਾਰ ਸਭ ਕੁਝ ਮੰਨਣ ਨੂੰ ਤਿਆਰ ਹਨ
1-M-19   **ਵ**ੋ**ਟ**ਾਂ ਅਕਾ**ਲ**ੀਆਂ ਨੂੰ ਪਾਉਣ ਦੀ ਸ਼ਰਤ  ਤੇ ਮੰਨ ਸਕਦੇ ਹਾਂ
1-M-20   **ਵ**ੋਟਰਾਂ **ਵ**ੱ**ਲ**ੋ **ਵ**ੀਂ ਉਨ੍ਹਾਂ ਨੂੰ **ਵ**ੋ**ਟ**ਾਂ ਪਾ ਕੇ ਜਿਤਾਉਣ ਦਾ ਭਰਵਾਂ ਹੁੰਗਾਰਾ ਮਿ**ਲ**  ਰਿਹਾ ਹੈ
1-M-21   **ਵ**ੋਟਰਾਂ ਨੇ ਆਪਣਾ ਫਰਜ਼ ਨਿਭਾ ਦਿੱਤਾ ਹੈ
1-M-22   **ਵ**ੋਟਰਾਂ ਨੇ ਆਪਣਾ ਕੰਮ ਨਿਭਾਅ ਦਿੱਤਾ ਹੈ
1-M-23   **ਵ**ੋਟਰਾਂ ਨੇ ਆਪਣਾ ਕੰਮ ਕਰ ਦਿੱਤਾ ਹੈ ਹੁਣ **ਵ**ਾਰੀ ਸਟੀਫਨ ਹਾਰਪਰ ਦੀ ਹੈ
1-M-24   **ਵ**ੋਟਰਾਂ ਨੂੰ ਨਸ਼ਿਆਂ ਦੇ ਜ਼ੋਰ ਨਾ**ਲ**  ਖਰੀਦਿਆ ਜਾਂਦਾ ਹੈ
1-M-25   **ਵ**ੋਟਰਾਂ ਨੂੰ ਕਿਸੇ ਤਰ੍ਹਾਂ ਦਾ **ਲ**ੋਕ **ਲ**ੁਭਾਉ ਜਾਂ ਸ਼ਰਾਬ ਕਬਾਬ ਜਾਂ ਹੋਰ **ਲ**ਾਲਚ ਨਹੀਂ ਦਿੱਤਾ ਗਿਆ
```

7. Known Issues or Limitations

A. Speech to Text:

- Currently, the whisper model is getting trained only on 1 epoch due to GPU container overuse (as stated by the IT section). The model performance is expected to improve on training with a larger number of epochs.

B. Keyword and Semantic Search:

- Keyword search and Semantic search work fine with given keywords, although the models can always be trained more on Punjabi Text files for better output.
- GUI runs on the local system and is not deployed to protect sensitive data and maintain information security.

8. Future Scope

A. Speech to Text:

- To train the model better, we have created an Augmented dataset, by adding augmentations (such as loudness, high pass filter, increased speed and gaussian noise) to the concatenated LDCIL dataset. The future batches may train the model on that dataset to get better results.

- Audio processing (apart from basic noise removal and frequency change) on Police files to improve model accuracy.

B. Keyword and Semantic Search:

- Before running Keyword Search on the given text file, we can write a code to replace its words with approx matches (exact+fuzzy) from the following Punjabi Word Dataset (Alphabetical Dictionary):

  https://www.kaggle.com/datasets/gurpejsingh/punjabi-alphabetical-dictionary-dataset

It is freely available on kaggle, and will ensure that the text file (transcripts) is more robust and readable.

- Semantic Model can be fine tuned on Punjabi text files for more specific use.
- Selecting the audio file, and running Whisper model for speech to text (using weights of model fine-tuned on LDCIL dataset) can be included in GUI for end-to-end use.
- Developing a GUI with React instead of Python as it's generally preferred due to limitations on access of desired python libraries and the ease of React GUIs to be extended on mobile applications.
- Discuss the exact deliverables with the concerned authorities and try to implement possible changes/ add ons thereafter.

9. Contact

- Akarshi Roy Choudhury – wb.akarshi@gmail.com, 2021eeb1149@iitrpr.ac.in, 8283848186
  (Wav2Vec, Whisper, GPU)
- Ashwini Sahane - sahaneashwini281@gmail.com, 2021eeb1159@iitrpr.ac.in, 7977537982
  (Wav2Vec, Whisper, GPU)
- Bhumika – bhumikachaudhari678@gmail.com, 2021eeb1160@iitrpr.ac.in, 7020719258
  (Keyword Search, Semantic, GUI, GPU)
- Lokesh Jassal – lokeshjassal2004@gmail.com, 2021eeb1185@iitrpr.ac.in, 9872807101
  (Semantic, GPU Specification Analysis)

10. Codes and Dataset

Link to the codes and dataset:
https://github.com/bhumikapc/Punjabi-Speech-to-text-Analysis-BTP-