Mini-Project:Speech Recognition and Transcription with NLP Post-Processing

Arun Kumar (11038660)

Applied Data Science and Analytics

Data Analytics 2:Text Mining and Natural Language Processing (K-2379-4343)

Prof.Doc. Swati Chandna

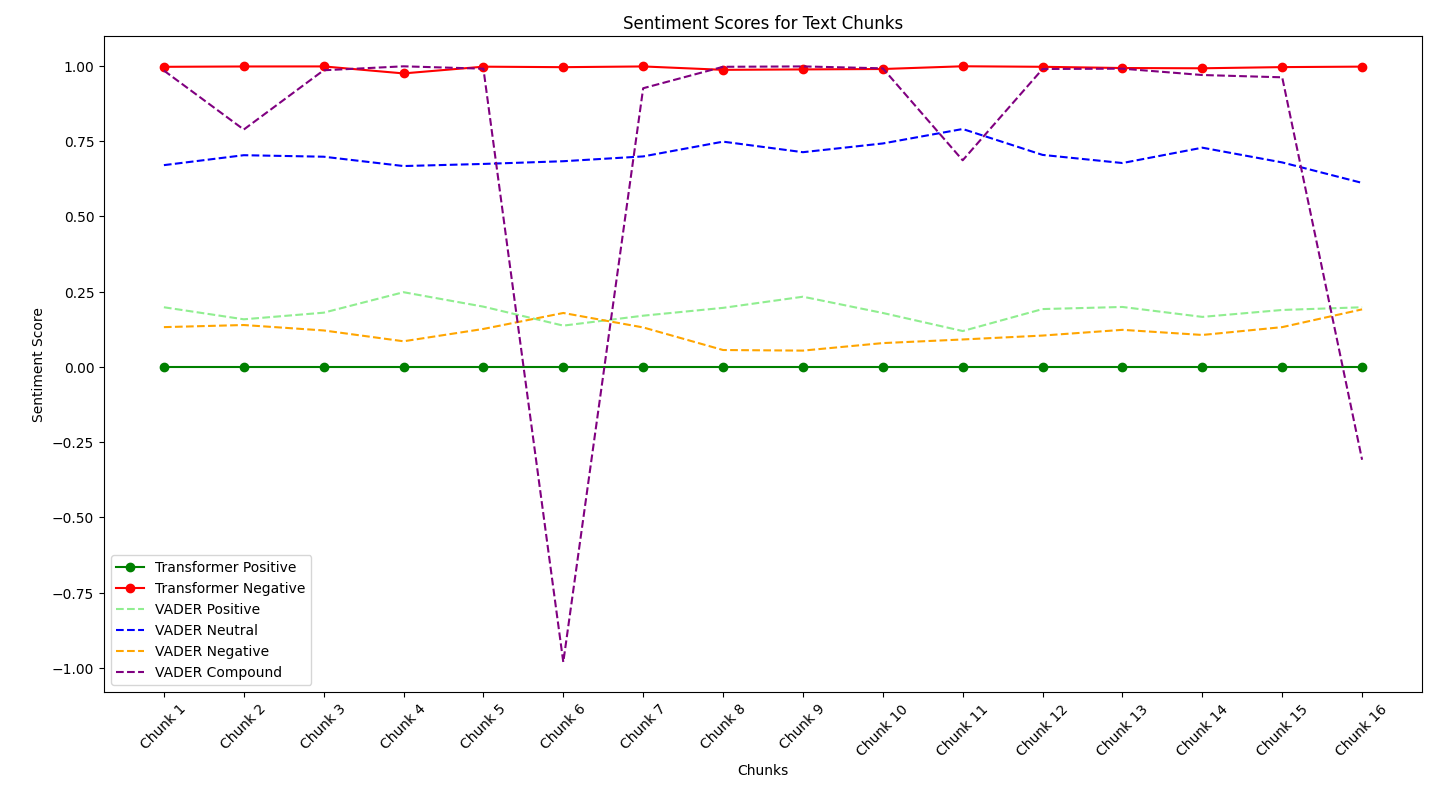
11 November 2024

**Goal:**

This project aims to leverage speech-to-text technology and Natural Language Processing (NLP) techniques to accurately transcribe audio files and extract valuable insights from the transcribed text.

**Methodology:**

1. **Speech-to-Text Transcription:**
   * **Mp3 to WAV File Conversion:** Librivox is used to download a free and open-source audio. This audio file is converted to wav format for ease of transcription.
   * **Audio to Text Conversion:** Assembly AI is used to convert the audio file to text and stored in a text file for further preprocessing of text. Assembly AI offers 50 USD worth free credits making it the best choice for this task.
   * **Text Preprocessing:** The text is passed through some steps such as lower case, removal of special characters, stop words and tokenization.
2. **NLP Post-Processing:**
   * **Sentiment Analysis:** Transformer based model and VADER are used for sentiment analysis. Chunking logic is also implemented to ensure not more than 512 tokens are sent to the model at once by leveraging the sliding window approach.
   * **Text Summarization:** The chosen text summarization approach utilizes a technique called TextRank. This method leverages sentence importance scores based on TF-IDF (Term Frequency-Inverse Document Frequency) and cosine similarity
   * **Topic Modelling:** Latent Dirichlet Allocation (LDA) is used for topic modelling as it's a probabilistic generative model that can discover abstract topics from a collection of documents, making it ideal for identifying the underlying themes in a set of transcriptions.



**Reasoning:**

* **Speech-to-Text API:** Assembly AI is used since it provides users with 50 USD worth free credit and ease of use.
* **Text Preprocessing:** Preprocessing is done to ensure normalization of data before pushing it to machine learning model such as lower case. Eg. ‘Ai’ and ‘ai’ should be considered the same. Stop words are used to reduce dimensionality and complexity. Tokenization is done to reduce computational costs and obtain a manageable structure.
* **Sentiment Analysis:** VADER is a sentiment analysis tool specifically designed for social media text, which often shares similarities with conversational language used in audio recordings. Its accuracy in detecting sentiment in such text makes it a suitable choice for this project.
* **Text Summarization:** By combining TF-IDF and cosine similarity, TextRank assigns a score to each sentence reflecting its importance within the context of the entire text. Sentences with high scores are considered more relevant and informative. The summary creation process then selects the top N sentences.
* **Topic Modelling:** LDA is a popular probabilistic generative model that can discover abstract topics from a collection of documents. It's well-suited for identifying the underlying themes in a set of transcriptions, even when the topics are not explicitly stated.

**Challenges Faced:**

1. **Data Challenges:**
   1. **Audio to Text Conversion:** 
      1. Difficulty finding a suitable free API with sufficient limits for large audio files.
      2. Issues with converting audio formats (e.g., MP3 to WAV).
   2. **Text Length Limitations:** 
      1. Transformer models have limitations on input text length (e.g., 524 tokens).
      2. Implementing effective chunking strategies to handle long texts.
2. **Processing Challenges:**
   1. **Text Preprocessing:** 
      1. Deciding when to use tokenization and lemmatization for optimal results.
      2. Balancing the trade-offs between accuracy and computational efficiency.
3. **Model Selection and Tuning:**
   1. Selecting the appropriate NLP models and techniques for specific tasks.
   2. Fine-tuning models to achieve desired performance on the given data.

**Conclusion:**

By combining speech-to-text technology and NLP techniques, this project aims to automate the process of audio analysis, enabling efficient extraction of valuable insights from audio content.

**References:**

1. NLP Libraries
2. Hugging Face
3. AssemblyAI
4. Machine Learning Libraries
5. Medium