Enhancing COVID-19 Data Visualization Through Large Language Models

Isaac Mylabathula, UNC Charlotte
Dr. Aidong Lu, Kexin Ding, College of Computing and Informatics

UNIVERSITY OF NORTH CAROLINA CHARLOTTE | COLLEGE OF COMPUTING AND INFORMATICS

Introduction

Addressing the complexities inherent in COVID-19 epidemiological data, this project employs advanced **Large Language Models (LLMs)** to refine data visualization and interpretation processes. Driven by the critical necessity for precise and accessible communication of health data, we have engineered tools that adeptly transform detailed datasets into comprehensible visual formats. Preliminary evaluations reveal enhanced usability; however, challenges persist with model-generated inaccuracies, necessitating further refinement to ensure both reliability and efficacy in real-world applications.

Large Language Models (LLMs): deep learning models pre-trained on extensive data

Methodology Continued

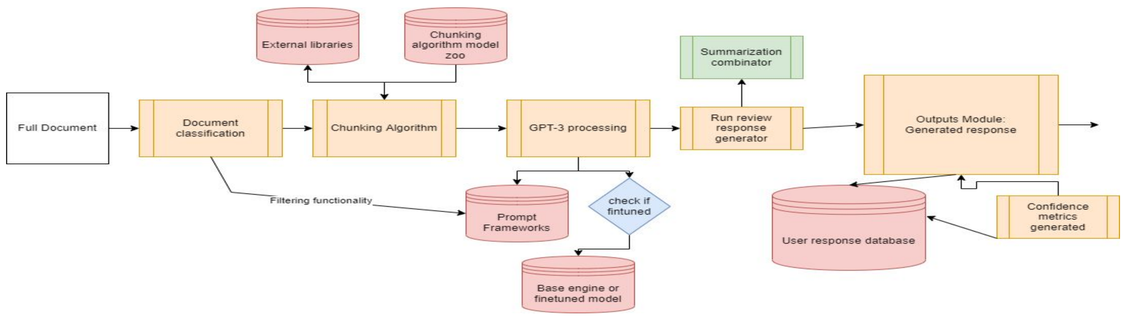

Figure 1: Text Summarization

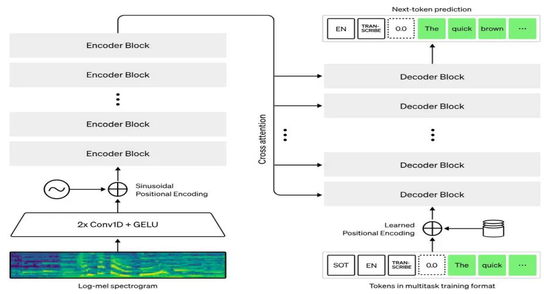

Figure 2: Audio-to-Text Conversion

Conclusions

The project findings underscore the potential of advanced data visualization tools to significantly enhance user understanding of complex epidemiological data. The use of LLMs for summarizing and analyzing academic literature and user feedback has proven effective, albeit with areas needing refinement, particularly in reducing hallucinations in model outputs.

Future Plans

- **Model Refinement:** refining the accuracy and detail orientation of LLMs, with an emphasis on eliminating hallucinations.
- **Tool Enhancement:** incorporating more dynamic and interactive elements into the visualization tools based on user feedback and summarized insights.
- **Broader Applications:** Exploring the extension of these tools and methods to other areas of big data analytics, potentially broadening the impact of this research.

References

Methodology

Data Collection

- **Academic Articles:** Comprehensive literature from IEEE and other databases provided insights into interactive visualization techniques and tools
- **User Interactions:** Data on user interaction with existing visualization tools.

Data Summarization

- Utilized LLMs to summarize content pertaining to interactive visualization techniques and tools

Data Conversion

- Converted audio data (user feedback) to tabular data using the Whisper machine learning model, followed by text clearing process to enhance clarity

Results 1

Data Summarization

- Users interacted with a tool that conducts visual analytics on COVID-19 epidemiological data.
- The tool visualizes frequent patterns using pie charts for 1-itemsets and sunburst diagrams for k-itemsets (k > 1).
- The visualization reveals that 90% of cases were transmitted through domestic acquisition, with 4% through international travel and 6% unstated transmission.
- Additionally, the tool displays that 48% of cases transmitted through domestic acquisition recovered without hospitalization.
- The visualization also shows hospitalization and death statistics, providing insights into the outcomes of COVID-19 cases.

Figure 3: Text Summarization via GPT-3.5

Your provided content discusses the use of a visual analytics tool applied to COVID-19 epidemiological data from Canada. This tool was utilized to examine case details, symptoms, clinical outcomes, and exposure methods for cases up to August 2020. It visualizes data through pie charts for single-item frequencies and sunburst diagrams for multiple-item frequencies. The findings show that the majority of transmissions were domestic, with a smaller portion attributed to international travel or unspecified. Additionally, the tool indicates that a significant percentage of those infected domestically recovered without hospitalization. A detailed breakdown of clinical outcomes for these cases is also presented, revealing the proportions of non-ICU hospitalizations, ICU admissions, recoveries without hospitalization, and fatalities. The visualization aids in understanding the extent of domestic transmission and its outcomes, including recovery and mortality rates.

Figure 4: Text Summarization via GPT-4

Results 2

Data Conversion

```
for file_name in audio_files:
    result = model.transcribe(os.path.join(audio_files_directory, file_name))
    cleaned_text = ' '.join(token.text for token in nlp(result['text']) if token.text.lower() not in ['yes', 'ok', 'good', 'next'])

    queries = [sent.text.replace('next', '').strip() for sent in nlp(cleaned_text).sents if "data" in sent.text.lower()]
    structured_queries = [f"{i}. {query}? 0.* for i, query in enumerate(queries, 1)]

    print(f"User: {file_name.split('.')[0]}")
    for query in structured_queries:
        print(query)
```

Figure 5: Audio-to-Text code

	A	B	C
1	User	Query	Value
2	User 7	1. , what is that current dot , the current data value account for ??	0
3	User 7	2. , what is that current data ??	0
4	User 7	3. That data value is at the more 0.8 .?	0
5	User 7	4. That data value of taxes ??	0
6	User 7	5. , what is the average data value of our telephone ??	0
7	User 7	6. , what is the average data value around 4 then ??	0

Figure 6: Overall Text Output

	A	B	C	D	E	F	G	H	I	J
1	id	seek	start	end	text	tokens	temp	avg_logprob	compression_ratio	no_speech_prob
2	0	0	0	7	Next, what is that current dot, the current data value account for?	[50364, 3087, 11, 437, 307, 300, 2190, 5893, 11, 264, 2190, 1412, 2158, 2696, 337, 30, 50714]	0	-0.7596588982	1.333333333	0.3436178267
3	1	0	7	9	Yeah.	[50714, 865, 13, 50814]	0	-0.7596588982	1.333333333	0.3436178267
4	2	0	12	15	0.3.	[50964, 1958, 13, 18, 13, 51114]	0	-0.7596588982	1.333333333	0.3436178267
5	3	0	15	17	Okay. Great.	[51114, 1033, 13, 3769, 13, 51214]	0	-0.7596588982	1.333333333	0.3436178267
6	4	0	17	21	Next, what is that current data?	[51214, 3087, 11, 437, 307, 300, 2190, 1412, 30, 51414]	0	-0.7596588982	1.333333333	0.3436178267
7	5	2100	21	28	That data value is at the more 0.8.	[50364, 663, 1412, 2158, 307, 412, 264, 544, 1958, 13, 23, 13, 50714] _	0	-0.5112097705	1.348214286	0.2146965265

Figure 7: In-depth Tabular Data