Assignment: Module Cal Capstone Project : Problem Statement (Module 17)
Project Name: Detect Comm
Student: Brad Brown

PROJECT QUESTION:
Using my previous research paper (title and link) as a starting point, can I improve on my initial research to achieve higher accuracy and more granular **detection of commercials within a television sports broadcast?**

The previous research paper was completed in March of 2024. The project fed 15-second “blocks” of text of 722 minutes of transcripts of the audio of TV sports broadcasts to ChatGPT via a python api, instructing it essentially to categorize each block as a commercial or a sports broadcast. Those answers along with other statistics about each block were then used to train a logistic regression model to try to improve the accuracy of the initial predictions.

Raw accuracy of individual block predictions improved from 87% to 91%, but a more granular attempt to categorize the data had a balanced accuracy of only 37%. The logistic regression model did not improve the balanced accuracy. Importantly, the accuracy was at its worst at the key transitions from a large contiguous set of sport broadcast blocks to a

set of commercial blocks. This result reduced its potential for practical use: Consumers care about accuracy around these transition points from mainly primary content to mainly commercial content.

The goal of the Cal Capstone project is to improve the prediction accuracy, especially at these moments when the television sports broadcast main content shifts to a set of television commercials.

Specific questions and goals

If the project de-emphasized the 15-second blocks of text concept and instead relied mainly on full sentences, would the accuracy and usefulness improve?

Assuming the evaluation of a single sentence will not give enough context to the LLM to categorize, can we create a mechanism to provide varying levels of context in addition to the sentence in multiple calls to the LLM and then use a voting mechanism to choose the final categorization? For example, will asking the LLM about the sentence and then asking the LLM again about the set of that sentence plus two prior sentences, help improve accuracy?

Can we train an XGBoost Decision Tree model on the LLM answers and metadata about each sentence to improve the accuracy more than the previous use of a Logistic Regression model?

Lastly, but importantly, should we conceive of and use different metrics for our results to align more closely with the business goals?

For example, should we consider a string of ‘flip flopping’ incorrect categorizations sentence by sentence worse than a consistent set of predictions that is slower to transition to the correct answer after a shift to or from commercials? In essence is it better to have your errors grouped together within long chains of sentences?

Also, can we look at a higher level metric that considers each conceptual commercial instead of categorizations of arbitrary blocks of text? How many specific commercials were correctly categorized completely, partially or not at all?

Similarly, should we differentiate between ‘embedded commercials’ vs ‘standard commercials’. For example, oftentimes a sports broadcaster will promote a product while still talking about the sporting event in the same sentence: “Let’s take a look at our Coca Cola top player of the game statistics…”.

Initial source of data:
Text transcripts of television sports broadcasts. The transcripts included every spoken word (not identifying the speaker). They included any spoken words from commercials as well. They were stored as 15-second elapsed time chunks of text in a custom SQL database schema.

My manual annotations captured one of 5 possible labels for each 15-second block:

- Sports Broadcast
- Sports Broadcast to Commercial
- Commercial
- Commercial to another Commercial
- Commercial to Sports Broadcast

Also for each block kept a simple binary Commercial Detected: Yes/No Label

A total of 2,891 15-second text blocks were captured in this initial project, representing ~722 minutes of video across 2 professional basketball games and 2 professional football games, one involving double over-time.

The already completed initial research project turned this transcript data into predictions for each 15-second text block. In addition, the ‘streaks’ of the presence (and sustained absence) of commercials and sports broadcasting blocks were calculated. Also,the final data of the initial research project incorporated the number of blocks of text of each type seen so far.

The data work involved would be to capture data in a schema that also captures the sentences within blocks and have predictions for each sentence.

Restructuring the data sets in this manner also requires handling the practical issue of “ground truth” labeling of each sentence. For example, in the previous research schema, a block could be labeled “Commercial-to-Sports Broadcast’. In the Cal Capstone project we need to label each sentence in that block as either a commercial or a Sports Broadcast. This is non-trivial considering there are 2,891 blocks in the data set. An efficient conversion human-in-the-loop workflow will be needed to convert the data.

Expected Results:

Overview:

- Improve raw accuracy higher from 91%
- Remove unneeded classes of output - simplify to Commercial: Yes/No
- Mechanism to optimize towards highest accuracy OR smoothest predictions (less likely to flip flopping over contiguous group of sentences)
- Define a metric to measure high-level accuracy in terms of end-user experience (beyond sentence level accuracy)

Single Sentence Focus:

The current 15-second blocks of text often have partial sentences at the beginning or end of the block. I expect working with complete sentences will improve the accuracy of predictions.

However, a sentence focus will introduce challenges. A single sentence may or may not provide the LLM enough context to decide its category. “Steph Curry makes a 3-point basket” will probably be very accurately categorized as a sports broadcast. Similarly, “Buy a Toyota today” will be a commercial. However, there are many cases where a single sentence will not be enough context, especially very short sentences like “Great!” or “It is a feeling we have before we make a move”. Therefore we need to add prior sentences to help the system understand the context. But adding context can also make the results worse. “Steph Curry makes a 3-point basket. Buy a Toyota today. Great! ” might still leave even a human wondering if the sentence “Great!” is part of a commercial or a sports broadcast. I believe these challenges can be

somewhat overcome by trying different levels of context and asking ChatGPT multiple times for a categorization. We can use heuristics to weight each response to make a final decision.

I do not know how much single sentences will improve the predictions but I am very confident the new structure will allow us to try a variety of modeling techniques and approaches quickly and easily.

Better Second Stage Model:
I expect the XGBoost model to provide better results than a logistic regression model. The traits of the XGBoost model lend themselves to this categorization problem. If that doesn't work, I will try a Neural Network model.

Different Metrics Than Previous Projects:
The previous project used 5 categories for each text block:

But three of these categories were because of the inherent use of elapsed time-based chunks of text. Multiple categories can be embedded within. By using sentences, it is probably far less likely that it will be both a commercial and part of sports broadcast. There are exceptions such as embedded sponsorships, such as the previously noted example: "Let's take a look at our Coca Cola top player of the game statistics…". However, I feel we can align these edge cases to the consumer's preference. Depending on the use case, most end-users would choose these cases as one or the other category and we can potentially configure the model accordingly.

Similarly, the previous research project and underlying system didn't give a good high-level picture of the business user's experience with the prediction system. Humans experience TV broadcasts at the conceptual level, not as arbitrary blocks of 15-second text. We should be able to understand how many commercials were detected by defining the beginning and end of each commercial.

Similarly, we should be aware of the context of a given prediction mistake. If there is a mistake in the middle of a key part of a sports broadcast, then that could have a big negative impact as opposed to making a mistake on the edge of a major transition from mainly sports broadcasting to mainly commercials. Let's assume the user would stop paying attention to

the sports broadcast if the system predicted a commercial. Doing so in the middle of the ‘action’ of a sports broadcast is probably much worse than missing the last few seconds before a sports broadcast breaks away to a set of commercials.

I expect a better metric system aligned to the data structure and to the user experience will improve the overall value of the project.

Expected Techniques:

NLTK to get to sentences:
Use the NLTK python package to convert 15-second text blocks to grammatically correct sentences.
The system will need to handle partial sentences at beginning and end of blocks by extracting from the previous block or subsequent block.

Database Schema changes:
Add a table to handle sentence predictions and link chunks to text block sentences.

Multiple LLM calls and voting mechanism:
Define a way to make multiple calls to chatGPT for the same sentence using different instruction prompts
Define a useful voting mechanism to produce a better stage 1 result.

Sliding window to get ‘windowed predictions’:
Sometimes errors will flip flop with correct predictions. For example, For 6 sentences, the predictions will be as follows:

- Sports Broadcast, Commercial, Sports Broadcast, Commercial, Sports Broadcast, Commercial

Will implement a way to choose to reduce this issue. Will try to use the prediction for the current sentence by giving the previous prediction(s) influence over the current prediction.

Annotation system changes:
Currently every transition was marked manually and the code will read that and associate the previous blocks and subsequent blocks appropriately as well as mark the block at that transition.
So if the annotation file says “Commercial-To-Sports Broadcast” at 3:50 seconds, it will:

1) Mark the block at 3:30 seconds a “Commercial”
2) Mark the block at 3:45 seconds a “Commercial-To-Sports Broadcast”
3) Mark the block at 4:00 seconds a “SportsBroadcast”
4) It will keep marking subsequent blocks as “Sports Broadcast” until it hits another transition in the annotation file entry

However, now we will need to modify this to mark each sentence in each block appropriately. So a sentence that starts before 3:50 would be a commercial and those starting after would be a sports broadcast. There would be no more “Commercial-To-Sports Broadcast” labels.

XGBoost modeling:
The second stage currently uses a Logistic Regression model. The features are:

- LLM Prediction (from Stage 1)
- Total Blocks seen so far
- Total Commercials seen so far
- Total Sports Broadcast seen so far
- Number of Commercials seen in a row in this streak
- Number of Sports Broadcasts seen in a row in this streak
- Number of blocks since seeing a commercial
- Number of blocks since seeing a Sports Broadcast

We will transition to train and fit an XGBoost model with the same features.

Wider Motivation of previous project and the Cal Capstone project:

Commercial detection in video content helps relevant groups analyze television and other video content. It helps advertisers confirm their content was properly aired as well as analyze the marketing content of their competitors. It also enables manufacturers of TVs and DVRs to create products that consumers want. There could be additional widespread value for other stakeholders.

Not directly related to commercial detection, advanced and active research is ongoing to enable automated detection of semantic topics and recognition of context. It is believed to be a key part of the much larger goal of AGI. Significant progress has been made in processing to allow systems to grasp higher level meaning of streaming chunks of text. Similarly, scene detection, object identification and spatial relationships are actively researched and developed in the image processing realm - many are trying to grasp the higher level context, intent and focus. Similar efforts in video processing are being done. Audio data while more mature is another area that is still being researched to enable AI tools to have better situational awareness.

Simultaneous to AI topic analysis research, for more than two decades, other researchers have proposed and analyzed automated commercial detection models. Their approaches range from black screen frame detection, high activity rate analysis, audio fingerprinting, color coherence vectors, aspect ratio change, scene transition detection, logo detection, and contrasting inputs versus databases of known commercials. They sometimes supplement using assumptions of time-constraints of commercials versus primary content. Linear regression analysis, SVMs, CNNs, all play a role but based on our search of the public research, none have published research about the use of large language models to detect commercials within television-based videos.

Next page is the APPENDIX which gives a more detailed description of the initial research project completed recently.

Helping an LLM improve its detection of TV Commercials in Sports Broadcasts

Brad Brown (bradb416@stanford.edu)

Department of Computer Science, Stanford University

Stanford
Computer Science

Project Overview

Can the topic analysis capabilities of large language models be used for television commercial detection?

- SCOPE: **Sports broadcasting on television** in the USA. The high-level goal was to **detect commercials** within short segments of the broadcast.
- PROCESS: A promising role for LLMs in this field, but **required additional 2nd stage modeling** techniques. No existing research has tried LLMs.
- RESULTS: Initial results detecting commercials given **15 seconds** of TV sports broadcast transcripts (text) gives a balanced **accuracy score of 79%, which were boosted to 87% via logistic classification.**
- ISSUES: 1) F**alse positives** 2) Previous detection methods using visual analysis reach 96% accuracy.

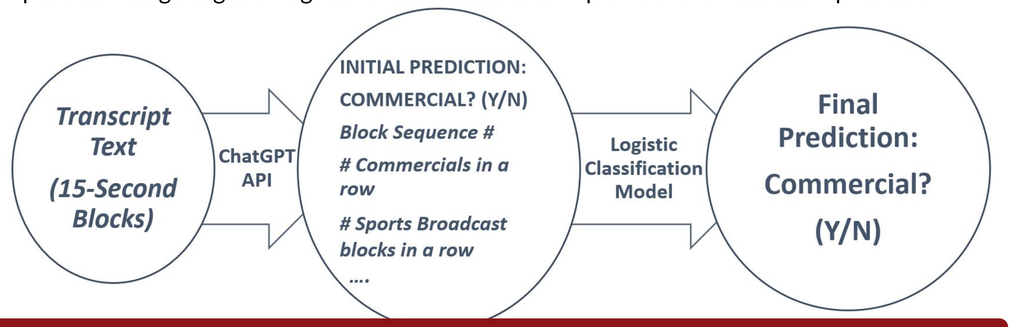
Datasets & Metrics

- LLM prompts and the transcript text blocks are the key input data. The transcripts included **every spoken word** (not identifying the speaker). They included any spoken words from commercials as well. There were stored as 15-second elapsed time chunks in a custom SQL database schema
- My manual annotations captured one of 5 possible labels for each 15-second block:
  - Sports Broadcast
  - Sports Broadcast to Commercial
  - Commercial
  - Commercial to another Commercial
  - Commercial to Sports Broadcast
- Also kept a simple binary **Commercial Detected:**
  - **Yes/No Label**
- A total 2,891 15-second text blocks were captured in this project, representing **~722 minutes of video** across 2 professional basketball games and 2 professional football games, one involving double over-time.

Methods & Experiments

The research work flow involved two high-level stages:

- Stage 1: The LLM prediction stage was to design LLM instruction prompts to classify chunks of text, feed those chunks to LLM API, and record results vs annotated labels.
- Stage 2: Final stage was to incorporate the LLM predictions, the 'streaks' of the presence (and sustained absence) of commercials and sports broadcasting blocks. Also, incorporated the number of blocks of text of each type seen so far. Together they formed the data input to post-LLM stage Logistic Regression SAGA model to improve on the initial LLM prediction.



Discussions & Future Research

Discussions:

- Logistic Classification Binary model measurably improved the accuracy - see "Classification report". Note the 4 and 8 point increases in accuracy and balanced accuracy to 91% and to 87%.
- The model with C = 1 regularization does not overfit ('Training Size vs Misclassification")
- False positives in commercial detection in both stages are a concern (~8%)
- The LLM Stage 1 and the LR Multinomial model (Stage 2) did NOT succeed in classifying transitions and combinations of events **within** the 15-second window. 37% Balanced Accuracy.
- No obvious game timeline pattern for accuracy of prediction - see "Commercials in a game"
- Insignificant differences between 2 experiment types: 15-second window/15-second context vs 15-second window/30-second context experiments.
- Best-performing LLM prompt instructed to predict categories, give its rationale for the answer, and make final second prediction within the same instruction, looking at its rationale.

Future Research:

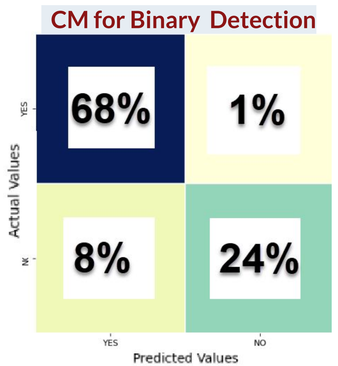
1) Systematically engineer and test additional prompts with full-scale testing vs a baseline
2) Capture the confidence or probability of LLM in its responses so I can better calibrate Stage 1
3) Try XGBoost and Poisson Point Process model as classifiers for the second stage
4) Develop an **ensemble approach with existing commercial detection techniques** such as black frame detection and volume variance
5) Dynamically change prompt to LLM as it is going through the event. Let it know what the second stage model thinks of its last prediction and also give it similar sequence data such as 'You predicted 4 commercial blocks in a row so far'. **Given the massive "memory" being added to LLMs, can give it the entire annotated training set from previous games. It is possible we don't need a second stage model.**

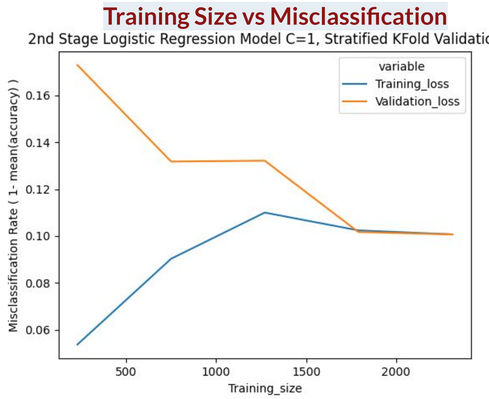
References

[1] R. Lienhart, C. Kuhmünch and W. Effelsberg. On the Detection and Recognition of Television Commercials. Universität Mannheim Praktische Informatik IV L15.16 D-68131 Mannheim

Results

Classification Report	Stage 1-->Stage 2 = Gain
Accuracy	0.87 --> 0.91 = +.04
Balanced Accuracy	0.79 --> 0.87 = +.08
Precision (macro avg)	0.91 --> 0.93 = +.02
Recall (macro avg)	0.79 --> 0.87 = +.08
F1-Score (macro avg)	0.82 --> 0.89 = +.07





91% accuracy with simple Log Reg model. Future research could yield even better results.