Artificial Intelligence COMPSCI 383 Group Project

Group Members: Samuel Warren, Ajay Manickham, Rikhav Shah

Problem Statement

Generating engaging and relevant commentary for American football plays is challenging due to the vast volume and complexity of play-by-play data. Sports analysts and fans seek real-time, entertaining insights during games, but existing methods often need more dynamism and contextual richness than a live commentary experience demands. This project aimed to address the gap by transforming structured play data into commentator-style commentary, ultimately enhancing the viewing experience for both casual fans and dedicated analysts.

Dataset

Our dataset consists of play-by-play football data sourced from Pro Football Reference. Each row in the dataset represents a unique play and contains attributes such as the quarter, game clock time, down, yards to go, field location, team scores, and a detailed textual description of the play. Unique plays are described as follows: "Brock Purdy pass complete short middle to Jauan Jennings for 13 yards (tackle by Rayshawn Jenkins)" Additional metrics include EPB (Expected Points Before) and EPA (Expected Points Added), which provide quantitative measures of a play's impact, but these were not used in creating the commentary, as their impact seemed irrelevant towards generating descriptive commentary for each play. We used web scraping tools like Requests to fetch HTML content and BeautifulSoup to create the dataset to parse the tables containing play-by-play data. The data was then collected, cleaned, organized, and grouped into a structured CSV file. A typical football game contains around 150–200 plays, meaning a full season across all teams can result in tens of thousands of entries. Unfortunately, our group instead focused on a dataset of around 500 entries or 3 games, since we have to evaluate each commentary manually, and reading thousands of entries would take a very long time. The dataset captures various play types, including runs, passes, kicks, timeouts, and penalties, with most plays occurring on 1st and 2nd downs. By combining information about the game situation and detailed textual descriptions, the dataset provides a comprehensive foundation for generating realistic, commentator-style football commentary.

Prompt Methodology

Prompt: "Give realistic, short, and concise commentary for this play. Use no outside context and just this information. Make sure all information in the commentary matches the information given. Respond only with the commentary. Quarter {quarter}, {time} time left, {down} down, {togo} yards to go at {location}. Score: {homeT}:{homeS}, {awayT}:{awayS}. Play: {play}"

Example Input: ,"Quarter:1, Time:15:00, Down:1, Yards to go:10, At:NYJ 30. Score: NYJ:0, IND:0. Play:Breece Hall right guard for 7 yards (tackle by E.J. Speed)"

Output: "Breece Hall takes the handoff and finds a gap through the right guard, gaining 7 yards before being brought down by E.J. Speed. It's a solid start for the Jets, setting up a more manageable second down. Score remains tied at 0."

We manually parsed the data.

Evaluation Approach

Our evaluation approach employed a Turing-like test, wherein the generated commentary was compared to similar radio commentary. Each commentary piece was rated on a scale from 0-7, with 0 indicating perfect commentary and higher values representing increasing errors, where 1, 2, and 3 correspond to one, two, and three errors, etc. The success rate was determined by calculating the percentage of commentary samples that received a rating of 0 and tracking the most common words used in the commentary for further analysis. This evaluation was conducted through a human-based method, allowing for nuanced judgment regarding the quality and realism of the generated commentary. The strengths of this approach lie in its potential for high accuracy, as human evaluators can incorporate subjective factors that contribute to the authenticity of commentary. However, a key weakness is the inherent human bias that may affect ratings, particularly regarding individual perceptions of what constitutes realistic commentary.

Results

```
Average Rating: 0.23732251521298176
Overall Success rate: 0.7971602434077079
Counts of each type of failure
Incorrect Down: 6
Incorrect Score: 51
Commentary was too wordy: 4
Incorrect Grammar: 3
Bad logic: 25
Incorrect Team: 28
Times they mentioned the score 418
Average Length 293.763
Mentions Exact Score:84.787% of the time
```

```
Most Used Words:
with: 547
down: 344
yards: 250
left: 238
their: 230
they: 196
this: 163
second: 158
short: 152
pass: 151
score: 137
just: 131
takes: 127
third: 122
fourth: 120
before: 109
crucial: 105
first: 105
handoff: 101
makes: 97
play: 96
lead: 96
right: 94
remaining: 93
```

```
connects: 90
face: 85
quarter: 85
drive: 82
solid: 82
remains: 80
gaining: 78
gain: 77
brought: 68
through: 66
being: 66
attempts: 66
facing: 65
trailing: 63
over: 62
from: 55
england: 54
deep: 53
strong: 53
detroit: 53
will: 50
still: 50
maye: 50
critical: 49
back: 48
delivers: 48
```

The evaluation data reveals a compelling assessment of the generated commentary's quality. With an average rating of approximately 0.24 and an overall success rate of 79.72%, the model demonstrated a high level of effectiveness despite its relatively small size. The counts of specific types of failures indicate areas for improvement; notably, incorrect scoring was the most frequent issue, occurring 51 times, followed by incorrect team mentions at 28 instances and logical discrepancies at 25. Additionally, the commentary was found to be excessively wordy on four occasions, and three grammatical errors were identified. Additionally, the model mentioned the exact score of 84.79% in its commentary, this is a very high percentage. To put this into perspective, that means that the exact score is mentioned in roughly 6 out of 7 plays. As a fan being reminded of the score this often might take away from the watchability of the game. Finally, the average commentary length was roughly 294 characters including spaces. If a typical word is 5 characters this leaves an average word count of around 50 when you factor out the spaces. This seems like a good sweet spot for length or radio-like commentary.

The second part of our commentary is a table representing all of the words, 4 letters or longer, with their counts. The table above represents the top 50 most used words. This allowed us to learn more about the word choice that chat-GPT uses for the commentary. Of the top 50 words, it is mostly comprised of prepositions such as with, right, or left, other common words in sentences such as this, that, they, etc., and football-related words; yard, down, pass, and score are all in the top 50. There is also a name and a team (Maye and Detroit) present likely because most of the data came from only a few games. Aside from these three groups, the commentary heavily relied on repetitive adjectives such as crucial, critical, and solid. This is important to note because to make the commentary seem more realistic, it would be better to limit the number of times each phrase is used. Simply changing or removing the adjective could improve a user's experience. Based on our group's personal experience, reading the same type of typical words in commentary made the process very repetitive, and frustrating.

Overall, these metrics suggest that while there are areas for refinement, the model performs remarkably well, especially given its small size, making significant strides in generating coherent and contextually relevant sports commentary. One possible way to improve the results is by adjusting our prompts to include more structure for the commentary so that we can stop certain phrases from being said. This would help improve cases where the commentary fails to pass the Turing test while keeping the generated commentary concise.