Voicing the Game: Advanced Language Models for Automated Sports Commentary

Vikranth Reddy Chapaala[†], Venkat Sai Phanindra[‡]

Watson Graduate School of Management, Spears School of Business, Oklahoma State University, Stillwater, OK 74078

[†]*vikranth.reddy@okstate.edu*, [‡]*phanindra.anagam@okstate.edu*

1. Abstract:

In the traditional domain of sports broadcasting, the reliance on manual commentary has been a norm, involving human commentators to deliver real-time insights, analyses, and play-by-play descriptions of the game. Despite its wide acceptance, this method is fraught with challenges, including human error, inconsistency, and the inability to cover every facet of the game in a comprehensive manner. Our research introduces an innovative approach to this problem by leveraging advanced language models to automate the process of commentary generation. While our methodology and findings are universally applicable across various sports disciplines, we have chosen cricket as an illustrative example to demonstrate our approach towards automating commentary generation and leveraging advanced language models to revolutionize the process.

A notable aspect of our project involves conducting sentiment analysis on the ball-by-ball commentary data from the Indian Premier League (IPL) 2018, for both batsmen and bowlers. The objective is to explore the prospects of automated commentary generation and assess the past commentary's tendency towards neutrality or positivity in sentiment. This paper presents our findings and discusses the potential of Language Modelling to revolutionize sports broadcasting by enhancing accuracy, consistency, and comprehensiveness.

2. Introduction:

In sports, commentary acts as the lifeblood of live broadcasts, turning quiet visuals into lively, emotive stories that capture the hearts of audiences worldwide. It's not just about narrating events as they happen; adept commentators add layers of context, thrill, and a personal touch to the games, enhancing the viewing experience from mere watching to full-on engagement. This crucial function of commentary in sports broadcasting highlights not merely a requirement but a fervent desire for vibrant, real-time storytelling that keeps up with the game's intensity. Live sports online streaming is a fast-growing market, forecast to grow from USD $18.6B in 2021 to USD $93.1B by 2027, a compound annual growth rate of 24.64% from 2022 to 2030. Sports streaming platforms continually advance, employing diverse strategies to delight users and enrich viewer engagement. [1]

However, sticking to the traditional way of producing commentary—where human commentators are tasked with covering each game—brings its own set of challenges and constraints, especially when it comes to providing in-depth coverage across various sports and events at the same time. Amidst these dynamics, the role of commentators, regardless of their geographic origins, plays a crucial part in shaping the narrative and the audience perception [2]. The need for expert insight, coupled with the overwhelming number of matches, can overextend resources and limit the ability to scale live commentary features across platforms.

According to IBM, its team sourced data from almost 130 million documents to train the large language model for Wimbledon commentary [3]. This is where the power of advanced language model's steps in to transform sports commentary. With the rapid advancements in artificial intelligence and machine learning, language models have grown increasingly adept at crafting coherent, context-aware, and captivating stories. They emerge as a beacon of hope against the backdrop of limitations tied to human-made commentary, offering a scalable, cost-efficient way to furnish a rich, engaging narrative experience across a wide range of sports applications.

For applications that are currently missing live commentary features, bringing in language models opens exciting new ways to pull in audiences, making the user experience far more dynamic and interactive when following live events. It's like having the chance to watch a game with a friend who's not only super knowledgeable but also never skips a beat.

Additionally, for platforms that already have human commentators doing their thing, adding language models into the mix can really beef up their game. It ensures even the less popular matches get their fair share of the spotlight, maintaining consistent coverage across the board. This mix-and-match strategy lets human commentators shine where it matters most, focusing on the big games, while AI takes care of covering the wide array of other events, making sure every game gets its moment.

Large language models (LLMs) have revolutionized the capacity for generating nuanced, context-aware text, simulating a degree of understanding and creativity that was previously the exclusive domain of human intelligence. These models, trained on vast corpora of textual data, excel in producing coherent and relevant narratives across various domains, including sports commentary. This methodology, adaptable and versatile, is applicable to any sport, offering a scalable solution to real-time commentary generation.

A significant advantage of employing language models for automated sports commentary lies in the utilization of data that is already being captured in real time by applications providing sports scores and updates, such as ESPN, Fox Sports, and others. These platforms collect detailed live game data, encompassing all the parameters necessary for generating rich, context-aware commentary. This existing infrastructure of data collection and processing presents a unique opportunity to augment the current offerings of sports applications with minimal additional investment.

The integration process is designed to be seamless and efficient, requiring minimal adjustments to the current data handling workflows of sports applications. By tapping into the existing data stream and feeding it into the chosen language model, such as GPT-4, the system can dynamically generate commentary that is both relevant and engaging, thereby enriching the live sports viewing experience without necessitating significant infrastructural overhauls or financial outlays.

At its core, weaving advanced language models into sports commentary is nothing short of revolutionary, opening the gates to high-quality, engaging sports stories for everyone. This exploration sheds light on the incredible potential of these technologies to change the game of live sports broadcasting, diving into how they can boost user involvement, widen access, and completely transform how we experience sports in this digital era.

3. Methodology

3.1 Sentiment analysis

We've made use of the ball-by-ball commentary data from IPL 2018, acquired through the ESPN Cricinfo API. The dataset consists primarily of raw information, including overs, short text containing details like bowler, batsman, and ball result, as well as longer text containing the human commentary.

3.1.1 Data Preprocessing

The short text containing the details was divided into three columns: bowler name, batsman name, and ball result. It was noted that the long text containing commentary provides insights into how the bowler delivered the ball and how the batsman reacted to it. To conduct sentiment analysis separately for both batsman and bowler comments, we split the commentary column into bowler and batsman commentary sections.

It was noticed that in most of the commentary entries, the batsman is referred either by their name or separated by a comma between the bowler and batsman remarks. Therefore, to accomplish this split we established our splitting criterion as the batsman's name. Rows that do not mention the batsman's name were split based on a comma.

3.1.2 Dimensionality Reduction

Here, we only retained the most relevant rows that cut down on the size of the data significantly. The step streamlined any subsequent phases related to analysis and training models by considering only data directly related to generating meaningful commentary. We have only considered overs, batsman and bowler commentary for our purpose.

3.1.3 Stopwords and Spelling Corrections

We utilized the NLTK library to conduct spell checks and remove stopwords from the text data. This preprocessing step is essential for text analysis as it helps enhance the accuracy and reliability of our results by eliminating common words that carry little semantic meaning and correcting any spelling errors that could affect the interpretation of the text.

3.1.4 Sentiment Analysis Models

We will be using 2 algorithms VADER (Valence Aware Dictionary and Sentiment Reasoner) lexicon and rule based bag of words model and Roberta Pre-trained model and compare both the models in terms of how well they can estimate the sentiment.

VADER utilizes a pre-built lexicon with sentiment scores for words, enabling quick sentiment analysis of text by assessing the polarity and intensity of individual words. VADER evaluates text by considering the presence and intensity of positive and negative words independently.

RoBERTa, stands for "Robustly optimized BERT approach" utilizes transformers for learning contextual representations of words and sentences. It is trained on diverse datasets with longer training durations to capture a broad understanding of natural language.

3.2 Commentary generation using Language Models

3.2.1 Prerequisites: Customizing Parameters Based on the Sport

The initial step in tailoring commentary generation to a specific sport involves identifying and formulating a set of key parameters that capture the essence and dynamics of the game. These parameters serve as inputs to the language model, guiding it to generate contextually appropriate commentary. For instance, in cricket, essential parameters might include Over number, Bowler name, Batsman name, Bowler action, Bowler outcome, Batsman action, Batsmen outcome, Delivery outcome

Similarly, for football, relevant parameters could encompass Quarter, Time Remaining, Down, Yards to Go, Team on Offense, Team on Defense, Offensive Play Type, Defensive Play Type, Player with Ball, Target Player, Outcome, Yards Gained, Play Result

These parameters capture the critical elements of each sport, enabling the language model to generate informed and engaging commentary that reflects the unfolding action.

3.2.2 Choosing the Model

Given their extensive training on a broad spectrum of English language data, large language models such as Code Llama, Gemini, Claude 2, Mistral 8x7, and GPT-4 are well-equipped to undertake the task of commentary generation. These models' ability to understand context, generate text that aligns with the given parameters, and maintain coherence over extended narratives makes them particularly suited for this application.

Due to computational constraints, this research utilizes the GPT-4 API to integrate the sport-specific parameters and generate commentary. This choice allows for leveraging GPT-4's advanced capabilities while managing the resource requirements effectively.

To demonstrate the practical application of this methodology, cricket data was chosen, and the commentary generation system will be hosted on a website for illustration purposes. This platform will serve as a real-time showcase of the language model's ability to produce live sports commentary. Furthermore, to evaluate the effectiveness and authenticity of the AI-generated commentary, a comparative analysis will be conducted between commentary produced by human experts and that generated by the language model.

This comparative study aims to assess the language model's performance in terms of accuracy, engagement, and the overall viewer experience, providing insights into the potential of LLMs to augment or complement traditional sports broadcasting methodologies. Through this analysis, we hope to highlight the strengths and limitations of automated commentary, paving the way for future enhancements in the field.

4. Results

4.1 Sentiment Analysis results:

To grasp the distribution of word lengths in both batsman and bowler commentary, we generated a density plot to visualize the data.

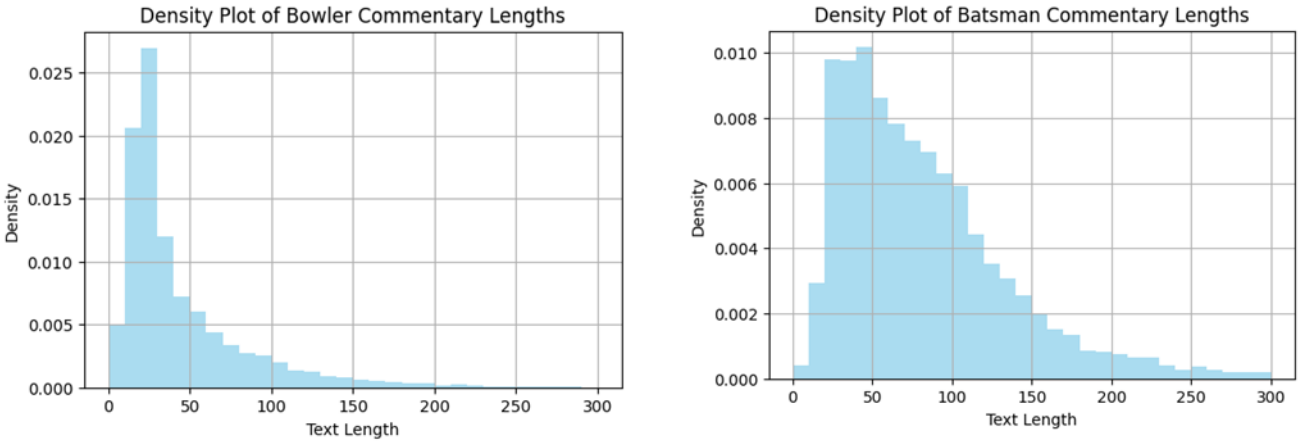


Fig 1. Distribution of bowler and batsman commentary lengths

We note that batsman commentaries tend to be longer in length compared to those of bowlers. Both distributions exhibit a right skew, with the majority of bowler commentaries falling within the 10-20 word range, while batsman commentaries are typically found within the 25-50 word range.

We have run VADER and Roberta models separately on batsman and bowler commentaries and here are the results.

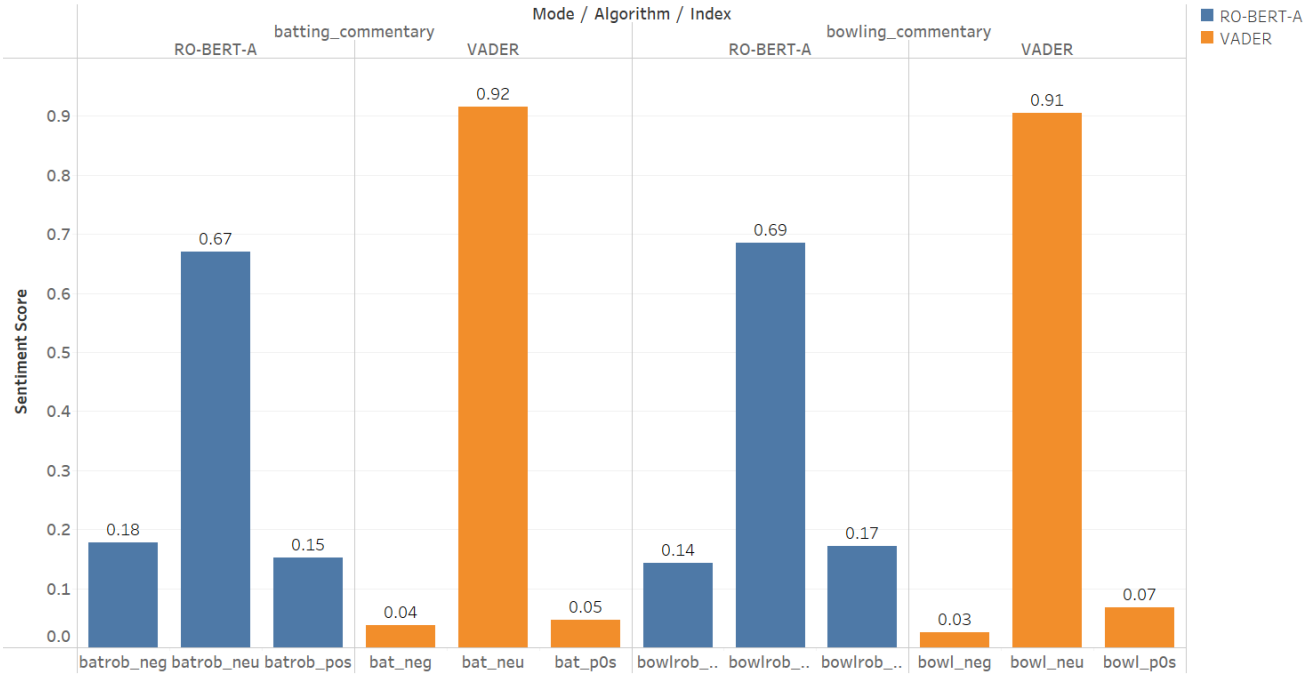


Fig 2. Sentiment scores of commentaries using both models

For the batting commentary:

The Roberta model scored the negative sentiment at 0.18, neutral sentiment at 0.67, and positive sentiment at 0.15, suggesting that the neutral sentiment is quite strong in batting commentary according to this model.

The VADER model scored the negative sentiment at a much lower 0.04, neutral sentiment at 0.92, and a positive sentiment at 0.05, indicating that there is neutrality in the batting commentary.

For the bowling commentary:

The Roberta model identified negative sentiment at 0.14, neutral sentiment at 0.69, and positive sentiment at 0.17. This distribution is somewhat similar to the batting commentary results from Roberta, with a majority of the sentiment being neutral.

The VADER model again shows a stronger skew towards neutral sentiment with negative sentiment at 0.03, neutral sentiment at 0.91, and positive sentiment at 0.07, similar to its results for the batting commentary.

The Roberta model is more sure of the sentiment that it predicts than the VADER model. This is because VADER does not consider the contextual use of words, rather it combines sentiments of individual words and provides a composite score . Conversely, the Roberta model incorporates the contextual meaning of words, resulting in more accurate sentiment assessments for the given text.

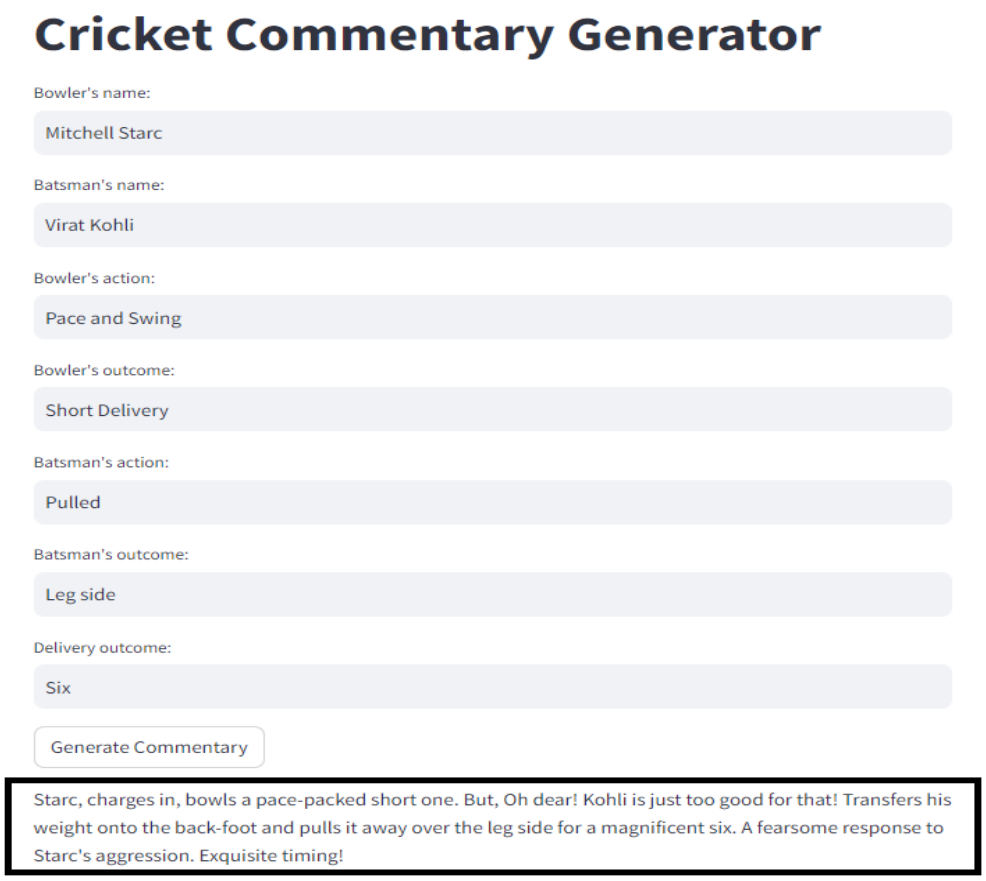
In batting commentary, the sentiment tends to be predominantly neutral, with a slight negative skew. Conversely, in bowling commentary, sentiment is largely neutral but slightly skewed towards the positive side.

4.2 Language Model results:

Based on the parameters discussed earlier for cricket, we have developed a custom application to demonstrate the outcomes of our code.

Before feeding the parameters to the model, I performed feature engineering on cricket commentary data from the IPL 2018 season, obtained via the ESPN Cricinfo API. The entire dataset comprises three columns: 'over', 'short_text', and 'long_text'. The 'short_text' column contains information about the batsman's name, bowler's name, and runs scored, while the 'long_text' column includes the human-generated commentary detailing the events of that delivery.

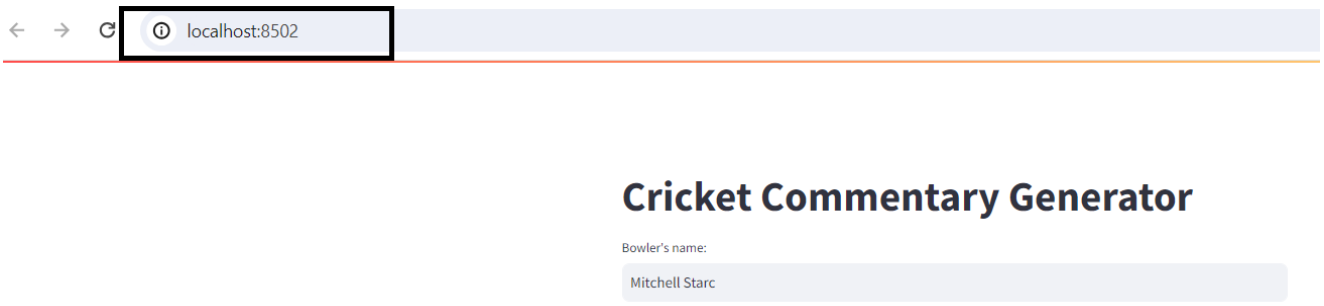
Using the 'short_text', I'm creating new columns for Batsmen, Bowler, and Runs. From the 'long_text', I am deriving additional features such as Delivery_Outcome, Bowler_Action, Batsman_Action, Bowler_Outcome, and Batsman_Outcome. All of this is accomplished using regular expressions and natural language processing techniques to parse and interpret the data efficiently.



In the above instance, we specified certain parameters for the model to generate commentary. Consequently, the model was able to produce the following:

"Starc, charges in, bowls a pace-packed short one. But, Oh dear! Kohli is just too good for that! Transfers his weight onto the back-foot and pulls it away over the leg side for a magnificent six. A fearsome response to Starc's aggression. Exquisite timing!"

The realism of the commentary generated based on the given parameters is evident, illustrating that the more detailed the parameters provided, the better the model performs across various game contexts. This effectiveness stems from the model having been extensively trained on a comprehensive dataset of commentaries from all matches to date.



This application is currently hosted locally for demonstration purposes but can be seamlessly integrated with existing systems to provide real-time commentary.

Now, let's examine how the commentary generated by our model surpasses human-generated commentary by analyzing a real-world example.

Parameters used:

over	Bowler	Batsman	Delivery_Outcome	Bowler_Action	Batsman_Action	Bowler_Outcome	Batsmen_Outcome
3.2	Watson	Rohit Sharma	no run	length, off	drive, pull	Not specified	Not specified
3.1	Watson	Ishan	1 run	bounce, length, off	backward point	outside off, length	third man

Result:

Over	Human Generated	Model Generated
3.2	Watson hits a hard length on off - neither drivable nor pullable - jabbed back to the bowler	Watson steams in, delivers a good length ball to Rohit. It's off the back foot, Sharma pulls but finds the fielder. No run there. A well-executed delivery by Watson and a guarded response from Sharma.
3.1	Back of a length and angling away outside off, Kishan hangs back, rides the bounce, and eases the ball to deep third man	Into the attack comes Watson, a short delivery outside off, Ishan quickly adjusts, guides it lightly to third man. They'll pick up a cool single.

Based on the profile and key achievements of the players involved, the model can also bring up relevant statistics. This feature enriches the commentary by providing historical context and comparative analysis, enhancing the overall viewer experience.

5. Conclusions

This research, universally applicable across all sports disciplines, showcases the transformative potential of advanced language models in automating sports commentary, with cricket serving as a demonstrative example. Through sentiment analysis of historical commentary and the generation of live, engaging narratives via LLM's like GPT-4, Llama, Gemini, Claude 2, Mistral, the study underscores the capacity of AI to enhance sports broadcasting significantly. It points to a future where AI not only complements human commentary but also expands the accessibility and depth of sports coverage, enriching the viewing experience for audiences worldwide. As these technologies evolve, the seamless integration of AI into sports commentary heralds a new era of innovation, making every game more engaging and accessible, irrespective of the sport in question.

6. References

1. https://aws.amazon.com/blogs/media/engage-online-sports-fans-with-live-event-commentary-using-generative-ai-on-amazon-bedrock/

2. Rawian, Rafizah & Khazin, Khairunnisa & Rullah Adha, T. Kasa & Siregar, Masitowarni & Sanjaya, Dedi. (2023). A Critical Stylistics Analysis of Sports Commentaries. World Journal of English Language. 14. 89. 10.5430/wjel.v14n1p89.

3. https://www.nbcnews.com/tech/innovation/us-open-ai-sports-commentary-masters-wimbledon-rcna95244