Predictive Text Analysis for Cricket Commentary

INFO 5731, Computational Methods, Section - 020 ; Submission Date: 03-10-2024

Sai Teja Rayabarapu
University of North Texas
Denton, Texas, USA ID: 11637122
SaiTejaRayabarapu@my.unt.edu

Manaswini Kodela
University of North Texas
Denton, Texas, USA ID: 11637117
ManaswiniKodela@my.unt.edu

1 INTRODUCTION

Cricket is a popular sport enjoyed by millions worldwide, and one of the most exciting aspects of cricket matches is the commentary that accompanies them. Traditional commentary is typically provided by human commentators who describe the action on the field, offer insights, and provide context to viewers. However, with advancements in natural language processing (NLP) and machine learning, there's an opportunity to explore generating cricket commentary automatically using algorithms.

The idea for this project arose from recognizing the potential of NLP and machine learning techniques in generating cricket commentary. By leveraging these technologies, we can automate the process of creating commentary, which could be valuable for various applications such as live match updates, video game simulations, or enhancing fan engagement on social media platforms.

Significance: Generating cricket commentary automatically has several significant implications. Firstly, it offers a novel approach to delivering real-time updates and insights to cricket fans, especially in situations where human commentators may not be available. Additionally, it provides a platform for experimenting with AI-generated content in the realm of sports commentary, potentially revolutionizing the way we consume and interact with sports media.

Research Questions: The primary research questions for this project include:

(1) Can we develop a model that accurately generates cricket commentary based on textual input?
(2) How does the quality of automatically generated commentary compare to that of human-generated commentary?
(3) What are the practical applications of AI-generated cricket commentary, and how can it enhance the viewer experience?

Research Purpose: The goal of this research is to develop a robust and accurate model for generating cricket commentary using NLP and machine learning techniques. By achieving this goal, we aim to provide cricket fans with an alternative source of commentary that is both informative and engaging.

Research Methods: To address the research questions and achieve the research goal, we will employ the following methods:

(1) Data collection: Gathering a diverse dataset of cricket match commentaries from various sources.
(2) Preprocessing: Cleaning and preparing the textual data for analysis, including tokenization, removal of stop words, and normalization.
(3) Model training: Training a neural network-based model, such as a recurrent neural network (RNN) or transformer architecture, on the preprocessed data.
(4) Evaluation: Assessing the quality of the generated commentary using metrics such as BLEU score, ROUGE, and human evaluation.
(5) Iterative refinement: Fine-tuning the model based on feedback and evaluation results to improve its performance.

Implications: The successful development of an AI-powered cricket commentary generator could have several practical implications. It could provide cricket fans with access to real-time updates and insights, enhance the viewer experience during live matches, and potentially reduce the dependency on human commentators in certain contexts. Moreover, it could pave the way for similar applications in other sports or domains where live commentary is essential.

2 RELATED WORK

Roy et al. [5] assert that in the realm of cricket analysis, the real-time commentary provided during matches stands out as a rich source of insights from seasoned players, offering genuine perspectives on unfolding events. Balaji et al. [1] further elaborate that this commentary, captured through microphone recordings and subsequent conversion to text, forms the basis for classification tasks aimed at understanding match dynamics. Rauf et al. [4]acknowledge the intricacies of predicting cricket match outcomes, influenced by the nuanced nature of player performances affected by opposition and venue dynamics. Hegde et al. [2] introduce innovative methods to overcome challenges in traditional live text commentary, leveraging dynamic web scraping techniques that gather real-time scores and parameters, feeding them into supervised learning algorithms to generate automated commentary. Wickramasinghe [10] highlights the challenges in sports data analytics, particularly in cricket, arising from the limitations of traditional statistical methods, necessitating alternative approaches to extract meaningful insights. Sinha [7] showcases the advent of machine learning in enhancing predictive capabilities, as seen in IPL match prediction models that integrate diverse factors like toss outcomes, home ground advantage, and player statistics through sophisticated algorithms. More et al. [3] emphasize the significance of text commentary, stating that it plays a pivotal role in conveying crucial match information, thus shaping the narrative surrounding live sports events. Srinivas et al. [8] highlight the meticulous process of analyzing cricket match data, which entails feature engineering, model selection, and performance evaluation of regression algorithms to ensure accuracy and reliability in forecasting outcomes. Sanjeeva et al. [6] demonstrate the role of speech recognition algorithms in facilitating the conversion of audio commentary into concise text summaries, enabling seamless integration into deep learning frameworks. Ul Abideen et al. [9] envision the future of cricket analysis with the development of automatic commentary generation models, leveraging state-of-the-art computer vision and natural language processing techniques to enhance the viewer experience and deepen insights into the game.

3 METHODOLOGY

Here are the research methods and techniques that are intended to be implemented for the proposed research questions

Research Question 1: Can we develop a model that accurately generates cricket commentary based on textual input?

Methodology and techniques: We will start by collecting a large dataset of cricket match commentaries from various sources such as live match updates, historical reports, and news articles. Next, we'll preprocess the data, cleaning it up and standardizing its format for analysis. Then, we'll train a neural network-based model, like a recurrent neural network (RNN) or transformer, using this preprocessed data. This model learns from patterns in the data to predict the next word or phrase in the commentary. We'll evaluate the model's performance using metrics like BLEU score and ROUGE. Finally, based on feedback and evaluation, we'll refine the model to improve its accuracy and coherence.

Research Question 2: How does the quality of automatically generated commentary compare to that of human-generated commentary?

Methodology and techniques: We'll compare the quality of automatically generated commentary to human-generated commentary using metrics like BLEU score and ROUGE, which measure the similarity between generated and reference text. Additionally, we'll conduct human evaluations where people judge the quality of both types of commentary. Through statistical analysis, we'll determine if there's a significant difference in quality between the two. This comparison will help us understand the strengths and limitations of AI-generated commentary compared to human-generated commentary.

Research Question 3: What are the practical applications of AI-generated cricket commentary, and how can it enhance the viewer experience?

Methodology and techniques: We'll explore practical applications of AI-generated cricket commentary by considering scenarios such as live match updates, video game simulations, and social media engagement. Through user studies and surveys, we'll gather feedback on the usefulness and engagement level of AI-generated commentary in these contexts. Additionally, we'll analyze social media interactions and viewer engagement metrics to assess the impact of AI-generated commentary on the viewer experience. This analysis will provide insights into how AI-generated commentary can enhance the sports viewing experience for cricket fans.

4 DATA COLLECTION AND CLEANING PLAN

4.1 Data Collection Strategy:

- We will start by identifying various sources of cricket match commentaries, including:
  - Live match updates from official cricket websites and sports news platforms.
  - Historical match reports from archives of cricket websites, sports news websites, and databases.
  - Cricket news articles and blog posts discussing match highlights and analysis.
- We will use web scraping techniques to extract text data from these sources, ensuring we gather a diverse range of commentaries spanning different matches, teams, and playing conditions.
- Additionally, we may explore the possibility of obtaining datasets from cricket databases or APIs provided by sports organizations.

4.2 Data Cleaning Process:

- Once the data is collected, we will proceed with the cleaning process to ensure it is suitable for analysis. The cleaning steps will include:
  - Removing irrelevant information such as advertisements, website navigation elements, and non-commentary text.
  - Standardizing the format of the data to ensure consistency across different sources.
  - Handling missing or incomplete data by either imputing missing values or removing instances with insufficient information.
  - Checking for and correcting any errors or inconsistencies in the text, such as misspellings or grammatical mistakes.
  - Removing duplicates to avoid redundancy and ensure each commentary instance is unique.
  - Filtering out any noise or irrelevant content that may not contribute to the analysis, such as irrelevant comments or unrelated discussions.
- We will also perform exploratory data analysis (EDA) to gain insights into the characteristics of the data and identify any additional cleaning or preprocessing steps required.
- Throughout the cleaning process, we will document each step taken and maintain a record of the original data and any transformations applied. This documentation will help ensure transparency and reproducibility of the cleaning process.

4.3 Quality Assurance:

- To ensure the quality of the cleaned dataset, we will conduct thorough quality assurance checks at various stages of the cleaning process.
- We will verify the correctness and integrity of the data by cross-referencing it with reliable sources and conducting manual inspections where necessary.
- Additionally, we will perform validation checks to ensure that the cleaned dataset meets predefined criteria and standards, such as data completeness, consistency, and accuracy.
- Any discrepancies or anomalies identified during quality assurance checks will be addressed promptly through appropriate corrective actions, such as revisiting the cleaning process or seeking expert advice.

- Finally, we will validate the final cleaned dataset by comparing it with known benchmarks or conducting pilot analyses to assess its suitability for the intended analysis tasks.

5 EXPERIMENT AND DATA ANALYSIS PLAN

5.1 Model Development Plan

- We will begin by splitting the cleaned dataset into training, validation, and testing sets. The training set will be used to train the model, the validation set will be used to tune hyperparameters and monitor performance during training, and the testing set will be used to evaluate the final model's performance.
- For model development, we will experiment with various neural network architectures, such as recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and transformer-based models like GPT (Generative Pretrained Transformer). These models have shown promise in generating coherent text and are suitable for sequence generation tasks like cricket commentary.
- We will train each model using the training data and optimize hyperparameters using the validation data to improve performance. Hyperparameters include parameters like learning rate, batch size, and model architecture-specific parameters.
- Throughout the model development process, we will monitor metrics such as loss function, perplexity, and validation scores to gauge the model's performance and make adjustments as needed.

5.2 Evaluation Plan

- Once the models are trained, we will evaluate their performance using the testing set. We will use both quantitative metrics and qualitative assessments to evaluate the quality of the generated cricket commentary.
- Quantitative metrics include BLEU score, ROUGE, and perplexity, which measure the similarity between the generated commentary and human-written commentary, as well as the coherence and fluency of the generated text.
- Qualitative assessments involve human judgment, where individuals will read and assess the quality of the generated commentary based on criteria such as relevance, informativeness, and readability.
- We will compare the performance of different models and select the one that achieves the best balance of quantitative metrics and qualitative assessments.

5.3 Data Analysis Plan

- In addition to evaluating the model's performance, we will also conduct exploratory data analysis (EDA) on the generated commentary to gain insights into its characteristics and identify any patterns or trends.
- EDA may involve analyzing the frequency of specific words or phrases, examining the distribution of sentiment or tone in the commentary, and identifying common topics or themes.
- We will use visualization techniques such as word clouds, histograms, and scatter plots to present the results of the data analysis in a clear and interpretable manner.
- The insights gained from the data analysis will help us understand the strengths and limitations of the model and identify areas for further improvement or refinement.

6 TABLES

The tabular representation below outlines the tasks and their respective timelines for the project. Each team member will be responsible for performing the tasks listed, as each person will select a dataset and execute all the operations described.

Table 1: Task assignment and timeline

Task	Team members	Timeline
Data Collection	Sai Teja Rayabarapu, Manaswini Kodela	Week 1
Data Cleaning	Sai Teja Rayabarapu, Manaswini Kodela	Week 2
Data Analysis and Visualization	Sai Teja Rayabarapu, Manaswini Kodela	Week 3
Comprehensive Literature Review	Sai Teja Rayabarapu, Manaswini Kodela	Week 4
Statistical metric analysis	Sai Teja Rayabarapu, Manaswini Kodela	Week 5, 6
Generative AI Analysis	Sai Teja Rayabarapu, Manaswini Kodela	Week 7, 8
Project Report	Sai Teja Rayabarapu, Manaswini Kodela	Week 9, 10
Project Presentation Slides	Sai Teja Rayabarapu, Manaswini Kodela	Week 10

REFERENCES

[1] A. S. Balaji, N. G. Vignan, D. S. Anudeep, Md. Tayyab, and K. S. Lakshmi. 2022. Cricket Commentary Classification. *Intelligent Data Communication Technologies and Internet of Things* (2022), 825–836. https://doi.org/10.1007/978-981-16-7610-9_60

[2] A. S. Hegde, K. Jha, S. Suganthi, and P. B. Honnavalli. 2021. Automating live cricket commentary using supervised learning. *Proceedings of Data Analytics and Management* (2021), 37–48. https://doi.org/10.1007/978-981-16-6285-0_4

[3] N. More, P. Fernandes, N. Bhuta, R. Suri, and A. Patil. 2022. Construction of sports articles using audio commentary. *ICT Systems and Sustainability* (2022), 611–618. https://doi.org/10.1007/978-981-16-5987-4_62

[4] M. A. Rauf, H. Ahmad, C. N. Faisal, S. Ahmad, and M. A. Habib. 2020. Extraction of strong and weak regions of cricket batsman through text-commentary analysis. In *2020 IEEE 23rd International Multitopic Conference (INMIC).* https://doi.org/10.1109/inmic50486.2020.9318089

[5] R. Roy, Md. R. Rahman, M. Shamim Kaiser, and M. S. Arefin. 2021. Developing a text mining framework to analyze cricket match commentary to select Best Players. *Lecture Notes on Data Engineering and Communications Technologies* (2021), 217–229. https://doi.org/10.1007/978-981-16-6636-0_18

[6] P. Sanjeeva, J. Ajith Varma, V. Sathvik, A. Abhinav Sai Ratan, and S. Mishra. 2023. Automated Cricket Score Prediction. *E3S Web of Conferences* 430 (2023), 01053. https://doi.org/10.1051/e3sconf/202343001053

[7] A. Sinha. 2020. Application of Machine Learning in Cricket and Predictive Analytics of IPL 2020. (2020). https://doi.org/10.20944/preprints202010.0436.v1

[8] S. Srinivas, N. N. Bhat, and M. Revanasiddappa. 2020. Data Analysis of cricket score prediction. (2020), 465–472. https://doi.org/10.1007/978-981-15-7234-0_42

[9] Z. Ul Abideen, S. Jabeen, S. Saleem, and M. U. Khan. 2021. Ball-by-ball cricket commentary generation using stateful sequence-to-sequence model. In *2021 International Conference on Communication Technologies (ComTech).* https://doi.org/10.1109/comtech52583.2021.9616676

[10] I. Wickramasinghe. 2022. Applications of machine learning in cricket: A systematic review. *Machine Learning with Applications* 10 (2022), 100435. https://doi.org/10.1016/j.mlwa.2022.100435