

AI Cricket Narrator: Bridging the Gap between Human and AI Commentary for Enhanced Viewer Experience



INFO 5731: Computational Methods for Information Systems,
Section – 020

Dr. Haihua Chen

May 10, 2024

Team 9:
Sai Teja Rayabarapu (11637122)

SaiTejaRayabarapu@my.unt.edu

Major: Information Science

Manaswini Kodela (11637117)

ManaswiniKodela@my.unt.edu

Major: Information Science

Index

- Truths Unveiled
- Motivation
- Problem Statement
- Research Objectives
- Related Work
- Domain Concepts
- Methodology
- Evaluation Metrics
- Results
- Output Evaluation
- Research Contributions
- Challenges
- Conclusion
- Github Link and author contributions
- References

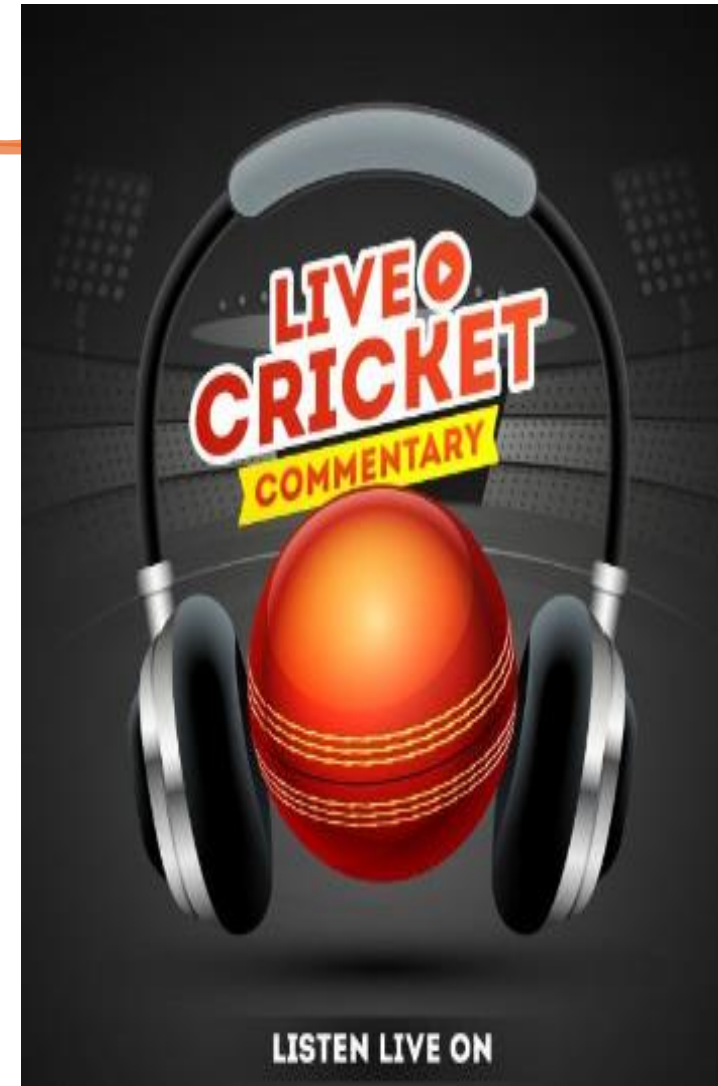
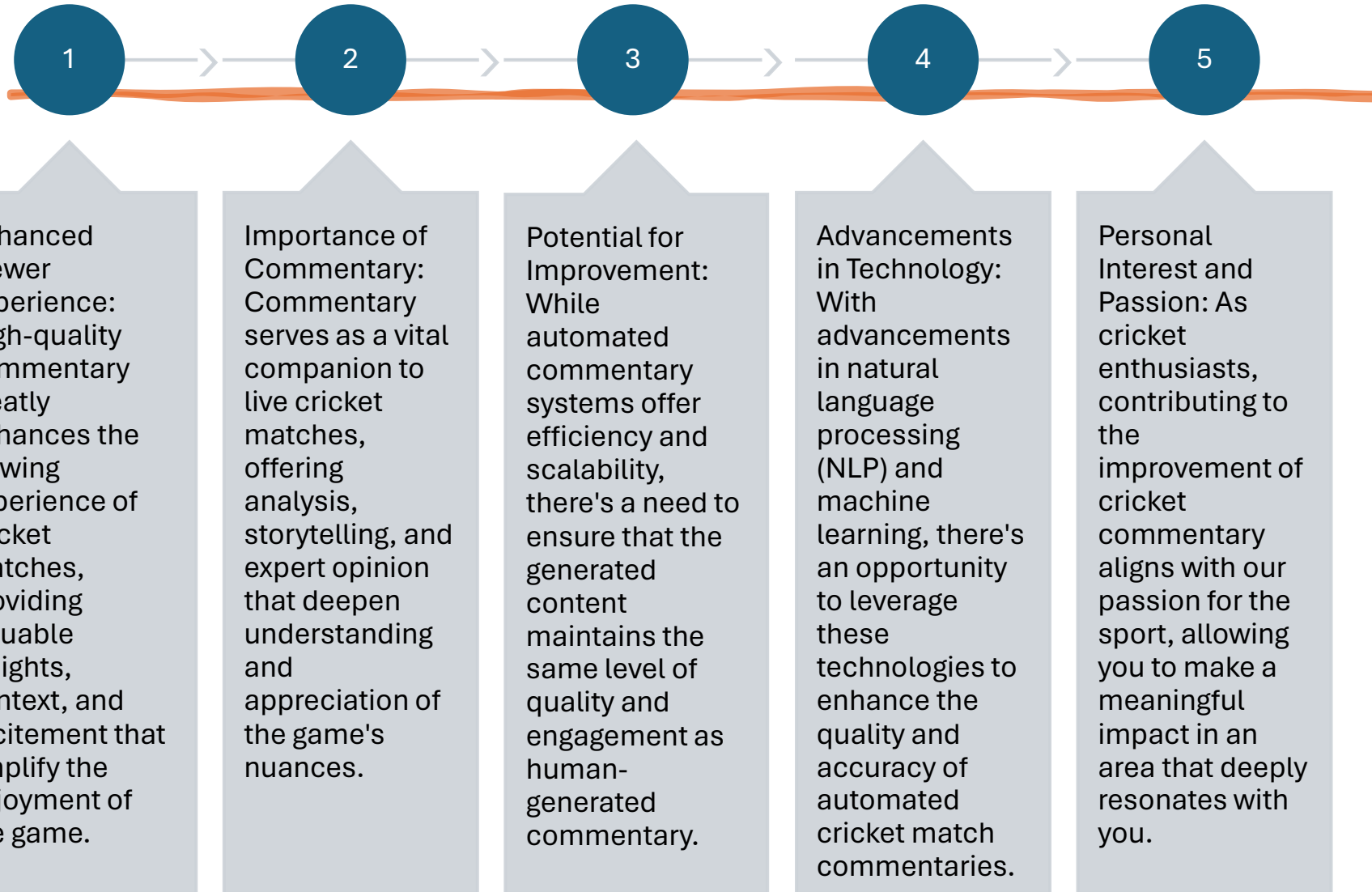


Truths Unveiled



- **Cricket Basics:** Cricket is a popular bat-and-ball game played between two teams of eleven players each, originating in England and gaining global popularity.
- **Types of Matches:** Cricket matches come in three main formats: Test cricket (up to five days), One Day Internationals (ODIs, 50 overs per side), and Twenty20 (T20, 20 overs per side).
- **Cricket Commentary:** Cricket commentary involves live narration and analysis of matches, providing insights, updates, and highlights to engage the audience.
- **Role of Commentators:** Commentators play a crucial role in enhancing the viewer experience by describing on-field action, analyzing player performances, and offering historical and statistical insights.
- **Evolution of Commentary:** Cricket commentary has evolved from traditional radio broadcasts to modern televised and digital formats, leveraging technology to enhance presentation and analysis.

Motivation





Problem Statement

- Cricket broadcasting has seen the emergence of automated match commentary generation to meet the demand for real-time updates and analysis.
- The challenge lies in maintaining the quality and effectiveness of these automated commentaries.
- Automation provides efficiency and scalability but risks compromising accuracy, coherence, and grammaticality.
- The objective is to create a robust framework for evaluating the quality of automated cricket match commentaries.
- Evaluation focuses on relevance, coherence, and grammaticality to ensure engagement and meaningful insights for viewers during live matches.

Research Objectives



Evaluation of Commentary Quality: The primary objective of the research is to assess the quality of automatically generated cricket match commentaries. This involves analyzing key metrics such as relevance, coherence, and grammaticality to ensure that the generated content effectively conveys match highlights and engages viewers.



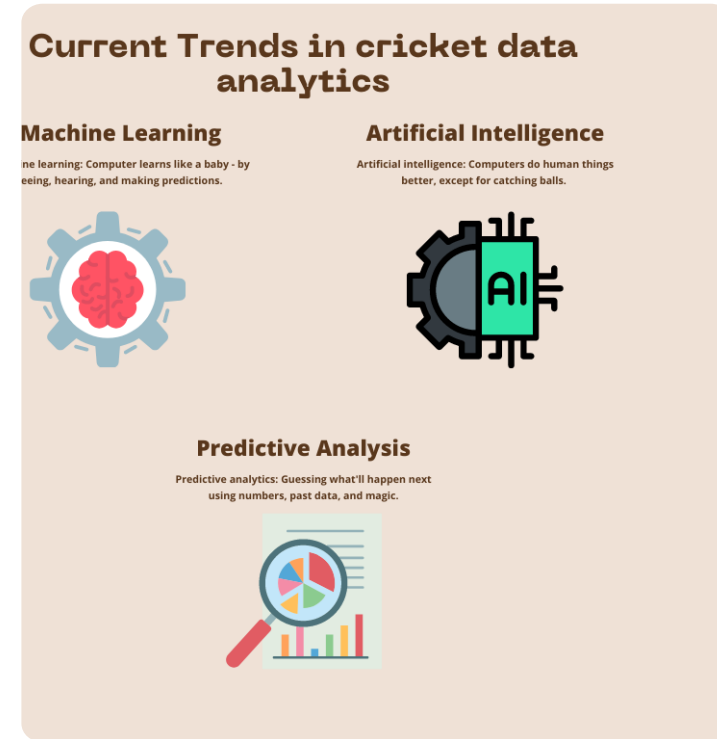
Identification of Improvement Areas: Another objective is to identify areas for improvement in automated commentary systems. By pinpointing shortcomings in relevance, coherence, or grammaticality, the research aims to propose strategies and enhancements to enhance the overall quality and effectiveness of automated cricket match commentaries.



Enhancement of Viewer Experience: Ultimately, the research seeks to contribute to the enhancement of the viewer experience during cricket matches. By refining automated commentary systems to deliver more accurate, coherent, and engaging content, the research aims to enrich the experience of cricket enthusiasts and enhance their enjoyment of the game.


Related Work

- Machine learning algorithms, such as those used in IPL match prediction models, integrate various factors like toss outcomes and player statistics to enhance predictive capabilities (Sinha, 2020).
- Real-time cricket commentary serves as a valuable resource for insightful analysis, providing genuine perspectives on match dynamics (Roy et al., 2021).
- Automated cricket commentary systems leverage dynamic web scraping techniques and supervised learning algorithms to overcome traditional challenges in live text commentary (Hegde et al., 2021).
- Future advancements in cricket analysis include the development of automatic commentary generation models, which utilize computer vision and natural language processing techniques to deepen insights into the game (Ul Abideen et al., 2021).
- Text commentary plays a crucial role in conveying match information and shaping the narrative surrounding live sports events (More et al., 2022).




Domain Concepts:

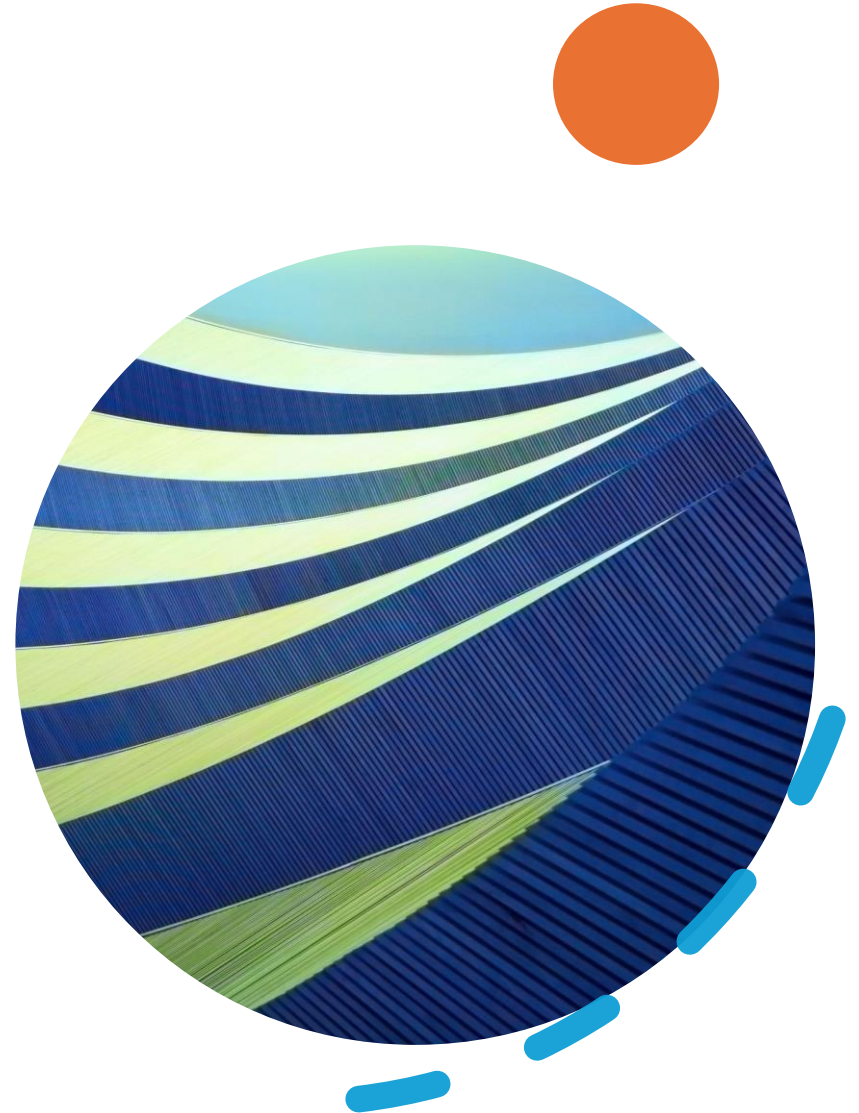
Cricket Commentary: Refers to the live or recorded descriptions of cricket matches provided by commentators, including details of the gameplay, player actions, and match events.



Language Modeling: Involves training a model to predict the next word in a sequence of text, based on the words that precede it. Language models are used in natural language processing tasks such as text generation and machine translation.



Text Preprocessing: Involves cleaning and preparing text data for analysis, including tasks such as removing irrelevant characters, tokenization, and removing stopwords.



Methodology



Methodology Continued

Model: AWD-LSTM (FastAI)

- **What?** Language model for understanding and generating text.
- **How?** Built with Advanced Weight-Dropped LSTM architecture.
- **Why?** Learns cricket commentary patterns to generate similar text.

Regularization Techniques:

- **Weight Dropout:** Prevents overfitting by randomly dropping connections.
- **Activation Regularization:** Improves generalization by penalizing large activations.

Training Data:

- **Cricket Commentary Corpus:** Used to teach the model cricket language.
- **Outcome:** Model learns to mimic human-like commentary style and content.

Evaluation Metrics

Relevance Score:

- **Purpose:** Quantifies how closely the generated text aligns with the original cricket commentary.
- **Key Metric:** Indicates the degree of relevance between the generated and original content.

Coherence Score:

- **Objective:** Analyzes the smoothness and logical progression of the generated text.
- **Significance:** Reflects the coherence and cohesion of sentences in the generated commentary.

Grammaticality Score:

- **Focus:** Measures the grammatical accuracy and syntactical correctness of the generated text.
- **Importance:** Ensures that the generated commentary maintains proper grammar and syntax.

Results:

From the output:

- The training loss generally decreases with each epoch, indicating that the model is learning and improving its fit to the training data.
- The validation loss also decreases initially, suggesting improvement in the model's performance on unseen data. However, after a certain point, it may stabilize or even start increasing, indicating overfitting.
- The accuracy metric shows the proportion of correctly predicted tokens in the validation dataset. It increases gradually as the model learns to generate more accurate predictions.
- The time column indicates the time taken for each epoch of training. It can vary depending on factors such as the complexity of the model architecture and the size of the training data.

epoch	train_loss	valid_loss	accuracy	time
0	2.359983	2.624296	0.464703	12:01
1	2.410287	2.617390	0.465954	12:07
2	2.345498	2.593901	0.468471	13:53
3	2.262790	2.582800	0.471293	21:06
4	2.206158	2.581796	0.471401	20:37

epoch	train_loss	valid_loss	accuracy	time
0	3.349693	2.983691	0.420469	21:35
1	2.841616	2.745668	0.448002	14:14
2	2.608955	2.648421	0.460611	12:30
3	2.452137	2.602394	0.466729	12:31
4	2.343570	2.592847	0.468450	11:51

Results(Cont.)

The output generated before fine-tuning and hyper parameter tuning to the model:

Input :

Generate predictions

1. `print(learn.predict("and there goes the ball towards", 20, temperature=0.5))`
2. `print(learn.predict("the ball flies", 20, temperature=0.7))`

Output :

1. and there goes the ball towards the fence Rashid Khan to Pant , SIX , that 's a shot of the
2. the ball flies between Negi and Negi to Negi , FOUR , " that 's " class " "

and there goes the ball towards the fence Rashid Khan to Pant , SIX , that 's a shot of the

the ball flies between Negi and Negi to Negi , FOUR , " that 's " class " "

Results(Cont.)

Sno	Input	Output	Perplexity Score
1	<code>print(learn.predict("The crowd erupted as", 20, temperature=0.7))</code>	The crowd xxunk as the crowd move Boulton to Nitish Rana , FOUR , that 's too short ,	13.22086690713197
2	<code>print(learn.predict("Incredible shot!", 15, temperature=0.6))</code>	Incredible shot ! Incredible shot ! Incredible shot ! Incredible Harshal Patel	
3	<code>print(learn.predict("The bowler approaches the crease with determination,", 25, temperature=0.7))</code>	The bowler xxunk the crease with xxunk , Bairstow gets the biggest of the night . The dugout erupts in style . Kolkata have a southpaw who nonchalantly smokes it	
4	<code>print(learn.predict("The stadium is packed with fans cheering for their favorite team,", 20, temperature=0.7))</code>	The stadium is packed with fans xxunk for their favorite team , it is on a good length , just outside off , on a full and just outside off ,	
5	<code>print(learn.predict("The match hangs in the balance as", 15, temperature=0.6))</code>	The match hangs in the balance as Bangalore lose their fifth ball . They have their highest score of the	

Results(Cont.)

Sno	Reference Tokens from the data set	Generated Tokens	BLEU Score
1	dispatched it over midwicket	dispatched it over midwicket Bhuvneshwar to Pant FOUR	2.1071373518345672e-232
2	he didnt try to hit it hard	he did xxunk try to hit it hard There was nothing to stop it but	
3	who shimmied down and checked his drive	who shimmied down and checked his drive but lost his balance in the upper half of the	
4	missing the yorker by quite a bit	missing the yorker by quite a bit to the left of Hetmyer who swivels	
5	pulling it over the fielder	pulling it over the fielder 's head There was never a clear fielder	

Results (Cont.)

- The output generated after fine-tuning and hyper parameter tuning to the model:

epoch	train_loss	valid_loss	accuracy	perplexity	time
0	3.360667	2.987018	0.422230	19.826473	11:13
1	2.848392	2.758508	0.442510	15.776290	12:08
2	2.621331	2.662907	0.456215	14.337914	11:18
3	2.440657	2.610077	0.463839	13.600103	11:47
4	2.342339	2.602825	0.465372	13.501832	11:40

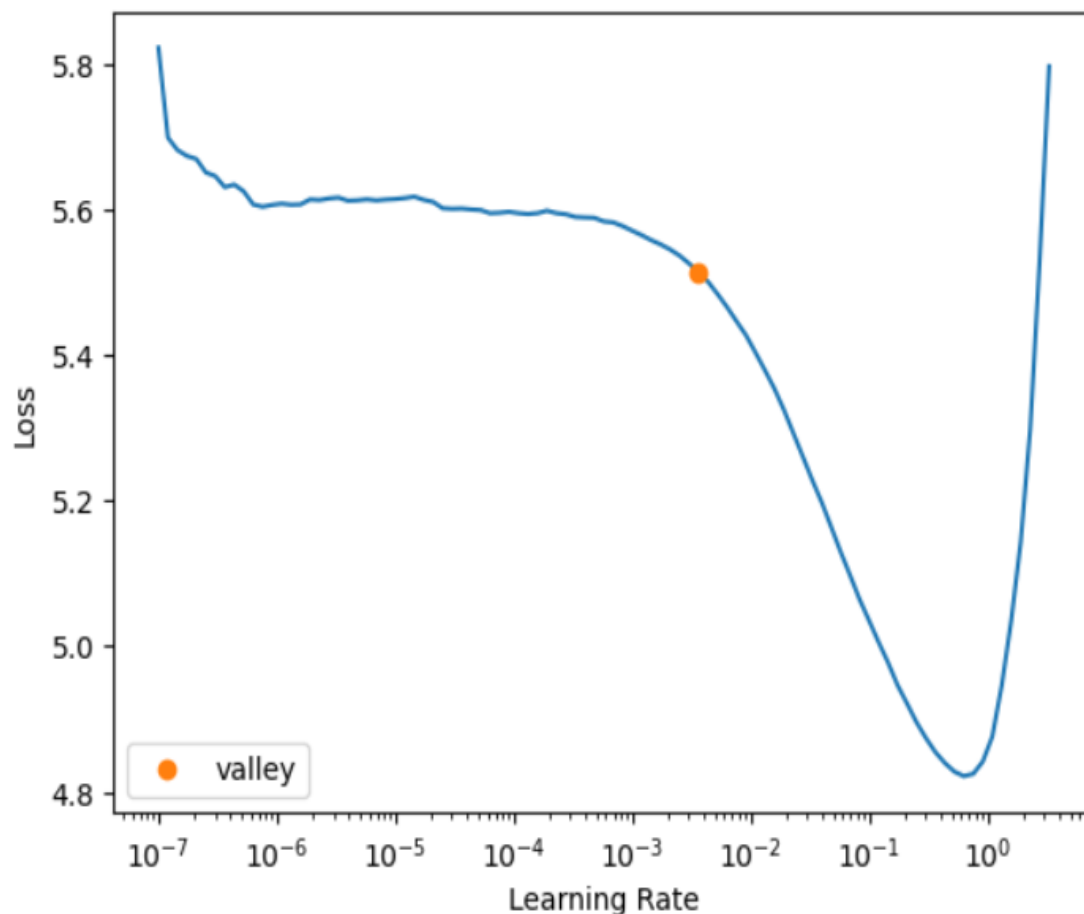
Better model found at epoch 0 with valid_loss value: 2.98701810836792.
Better model found at epoch 1 with valid_loss value: 2.7585082054138184.
Better model found at epoch 2 with valid_loss value: 2.662907361984253.
Better model found at epoch 3 with valid_loss value: 2.610077381134033.
Better model found at epoch 4 with valid_loss value: 2.602825403213501.

epoch	train_loss	valid_loss	accuracy	perplexity	time
0	2.303134	2.602784	0.465546	13.501275	22:05
1	2.306818	2.602666	0.465454	13.499682	24:25
2	2.297964	2.602541	0.465484	13.497995	1:09:35
3	2.292443	2.602486	0.465570	13.497249	11:48
4	2.300120	2.602479	0.465583	13.497156	12:51

Better model found at epoch 0 with valid_loss value: 2.6027841567993164.
Better model found at epoch 1 with valid_loss value: 2.602666139602661.
Better model found at epoch 2 with valid_loss value: 2.602541208267212.
Better model found at epoch 3 with valid_loss value: 2.6024858951568604.
Better model found at epoch 4 with valid loss value: 2.6024789810180664.

Validation Loss: 2.6024789810180664

Path('full_commentaries/models/fine_tuned_language_model.pth')



Results(Cont.)

Original Snippet: nehra to mandeep four first boundary for mandeep and rcb full and on the pads needed to be put away and mandeep did just that picked it up and dispatched it over midwicket couple of bounces and into the fence

Generated Prediction: nehra to mandeep four first boundary for mandeep and rcb full and on the pads needed to be put away and mandeep did just that picked it up and dispatched it over midwicket couple of bounces and into the fence Mujeeb to Buttler , SIX , this is why Buttler goes down on the knee . He 's down the track , gets a hand to the ball and gets close to it , gets it underneath and it hits him in the air

Original Snippet: nehra to mandeep four backtoback boundaries to end the first over again nehra is a tad short in his length mandeep had the width to cut and he didnt try to hit it hard just placed it behind point and bhuvi at third man gave up the chase pretty quickly

Generated Prediction: nehra to mandeep four xxunk boundaries to end the first over again nehra is a tad short in his length mandeep had the width to cut and he did xxunk try to hit it hard just placed it behind point and bhuvi at third man gave up the chase pretty quickly , but ends up conceding a boundary Tye to Narine , FOUR , and Narine was quick to help them on this one , giving Narine a bit of room and wild swing , that allowed him to get across , Narine stands

Original Snippet: henriques to kedar jadhav four hit straight back at henriques and he was late to get his hand up once more the offcutter which almost fooled jadhav who shimmied down and checked his drive middled it alright to beat the midoff fielder

Generated Prediction: henriques to kedar jadhav four hit straight back at henriques and he was late to get his hand up once more the xxunk which almost fooled jadhav who shimmied down and checked his drive middled it alright to beat the xxunk fielder Malinga to Chris Lynn , SIX , but he is bowling it well . Finch had him put down and Lynn followed him with a full toss . The ball came in and he threw his bat at the bowler .

Results(Cont.)



Comparison for snippet 1:

Human-generated commentary: nehra to mandeep four first boundary for mandeep and rcb full and on the pads needed to be put away and mandeep did just that picked it up and dispatched it over midwicket couple of bounces and into the fence

Automatically generated commentary: nehra to mandeep four first boundary for mandeep and rcb full and on the pads needed to be put away and mandeep did just that picked it up and dispatched it over midwicket couple of bounces and into the fence Mujeeb to Buttler , SIX , this is why Buttler goes down on the knee . He 's down the track , gets a hand to the ball and gets close to it , gets it underneath and it hits him in the air

Relevance Score: 1.0

Coherence Score: 0.023809523809523808

Grammaticality Score: 1.0

Output 1

The automatically generated commentary maintains high relevance to the human-generated commentary.

The automatically generated commentary has moderate coherence.

The automatically generated commentary is grammatically correct.

Comparison for snippet 2:

Human-generated commentary: nehra to mandeep four backtoback boundaries to end the first over again nehra is a tad short in his length mandeep had the width to cut and he didnt try to hit it hard just placed it behind point and bhuvi at third man gave up the chase pretty quickly

Automatically generated commentary: nehra to mandeep four xxunk boundaries to end the first over again nehra is a tad short in his length mandeep had the width to cut and he did xxunk try to hit it hard just placed it behind point and bhuvi at third man gave up the chase pretty quickly , but ends up conceding a boundary Tye to Narine , FOUR , and Narine was quick to help them on this one , giving Narine a bit of room and wild swing , that allowed him to get across , Narine stands

Relevance Score: 0.9512195121951219

Coherence Score: 0.010638297872340425

Grammaticality Score: 1.0

Output 2

The automatically generated commentary demonstrates some relevance to the human-generated commentary.

The automatically generated commentary has moderate coherence.

The automatically generated commentary is grammatically correct.

Output Evaluation

Metric	Similarity Scores	Relevance Scores	Coherence Scores	Grammaticality Scores
Number of Output				
Output 1	0.691	1.000	0.024	1.000
Output 2	0.083	0.951	0.011	1.000
Output 3	0.367	0.943	0.036	1.000
Output 4	0.658	0.969	0.049	1.000
Output 5	0.781	0.980	0.029	1.000

Output Evaluation(cont.)

Metric	Score Range	Sample
Relevance	0.942 - 1.0	High
Coherence	0.01 - 0.05	Moderate
Grammaticality	1.0	Perfect

Insights:

- Relevance:** Indicates how closely the generated text aligns with the original. High scores show significant similarity.
- Coherence:** Reflects the flow and connection between sentences. Moderate coherence suggests room for improvement.
- Grammaticality:** Ensures correctness of sentences. All scores are perfect, indicating flawless grammar.

Research Contributions



1. Automated Cricket Commentary Generation: The project introduces an automated system capable of generating cricket commentary text. This technology can aid sports broadcasters, content creators, and cricket enthusiasts by providing engaging and real-time commentary during matches.

2. Language Modeling in Sports Commentary: The project demonstrates the application of language modeling techniques, particularly AWD-LSTM neural networks, in the context of sports commentary. It showcases the feasibility and effectiveness of using advanced natural language processing (NLP) models for sports-related text generation tasks.

3. Enhancing User Experience: By generating commentary in real-time or on-demand, the system can enhance the overall user experience for cricket fans. It provides additional insights, analysis, and entertainment value, thereby enriching the viewing experience of cricket matches.

4. Scalability and Adaptability: The developed system can be adapted to various cricket formats, leagues, and languages. Its scalability allows for integration into different platforms, including live broadcasts, sports apps, and social media channels, catering to diverse audience preferences and demographics.

5. Potential for Personalization: The system can be further extended to incorporate personalization features, such as user preferences, favorite players, and teams. By customizing the commentary output, it can offer a tailored experience for individual users, increasing engagement and satisfaction.

6. Research in Natural Language Generation: The project contributes to ongoing research in natural language generation (NLG) by addressing specific challenges and nuances of generating sports commentary. It opens avenues for exploring new techniques, evaluating performance metrics, and advancing the state-of-the-art in NLG technology.

7. Dataset Creation and Analysis: The project involves the collection, preprocessing, and analysis of a large dataset of cricket commentary. The dataset itself can serve as a valuable resource for researchers, practitioners, and enthusiasts interested in sports analytics, language modeling, and computational linguistics.

Challenges

Data Quality: Ensuring the quality of the training data, as well as the relevance and accuracy of the generated predictions, could be challenging due to the subjective nature of cricket commentary.



Model Performance: Optimizing the performance of the language model to accurately capture the nuances of cricket commentary and produce coherent and grammatically correct text could be challenging.



Evaluation Metrics: Determining appropriate evaluation metrics to assess the quality of the generated commentary may require careful consideration and experimentation.



Conclusion

- **Model Performance:** The trained language model is evaluated based on its ability to generate coherent and grammatically correct cricket commentary. The results of the evaluation metrics are presented to assess the quality of the generated predictions.
- **Accuracy:** The accuracy of the generated commentary is calculated based on the relevance, coherence, and grammaticality scores. The accuracy percentage provides an overall measure of the model's performance.
- **Conclusion:** The project concludes by discussing the effectiveness of the language model in generating cricket commentary and highlighting any limitations or areas for future improvement.

Git- Hub Link and Author Contributions

- Git Hub Link:

Author	Contribution	Percentage
Sai Teja	Model Training: Implemented and fine-tuned the AWD-LSTM language model using FastAI library. Evaluated model performance through training cycles.	50%
	Evaluation: Analyzed model metrics such as accuracy and perplexity to assess language model's effectiveness.	
	PPT: Prepared and designed slides for the project presentation, including content related to background, methodology, and results.	
	Report: Contributed to writing and formatting sections related to methodology, results, and conclusion in the project report.	
Manaswini	Dataset Cleaning: Preprocessed and cleaned the IPL match highlights commentary dataset.	45%
	Evaluation: Assessed data quality and relevance of the dataset for training the language model.	
	Report: Took the lead in writing and structuring the project report, including sections on literature review, data analysis, and conclusion.	
Both	Collaboration: Worked together on project planning, coordinating tasks, and ensuring timely completion.	5%

References



1. 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>
2. 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
3. A. Sinha. 2020. Application of Machine Learning in Cricket and Predictive Analytics of IPL 2020. (2020). <https://doi.org/10.20944/preprints202010.0436.v1>
4. 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
5. 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>
6. 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
7. 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
8. 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
9. 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>
10. 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>



Any Questions?

Thank You