# AI Generated Cricket Score using NLP

Ms. Pranali Wagh
*Department of Information Technology*
*SAKEC*
Mumbai, India
pranali.wagh@sakec.ac.in

Sahil Desai
*Department of Information Technology*
*SAKEC*
Mumbai, India
sahil.desai16173@sakec.ac.in

Purav Doshi
*Department of Information Technology*
*SAKEC*
Mumbai, India
purav.doshi16174@sakec.ac.in

Chaitanya Gajoor
*Department of Information Technology*
*SAKEC*
Mumbai, India
chaitanya.gajoor16530@sakec.ac.in

Advait Narkar
*Department of Information Technology*
*SAKEC*
Mumbai, India
advait.narkar16652@sakec.ac.in

***Abstract*—To create an AI-based system for creating real- time cricket scorecards using live or recorded commentary, this study investigates the integration of Natural Language Processing (NLP), audio recognition, and machine learning approaches. Along with team names and venue information, users can stream live commentary using a microphone or upload audio files using the system's user-friendly frontend, Streamlit. Speech recognition is used to process and turn the audio into text, which is subsequently tokenized and subjected to NLP techniques to extract important events like runs, wickets, and overs. The scorecard is updated continuously by appending this textual data to an already-existing match commentary file. Additionally, a T20 dataset is used to train a Random Forest-based machine learning model that uses the dynamically generated scorecard data to predict the final match score. By providing both live updates and predicted insights, the system seeks to improve user experience by delivering an automated, real-time cricket score creation tool.**

***Index Terms*—Natural language processing (NLP), speech recognition, artificial intelligence (AI), cricket scorecards, real-time commentary, machine learning, random forests, Text analysis, tokenization, predictive modeling, Sports Data Extraction, Audio to Text Conversion, Automated Sports Analytics, and T20 Cricket Dataset**

## I. Introduction

Numerous industries have been transformed by the use of artificial intelligence (AI) and natural language processing (NLP), and the sports industry is no exception. Automating and improving live sports commentary analysis is one area where AI's potential is being investigated more and more. This study offers a fresh method for producing cricket scores in real-time through AI-driven approaches, particularly by fusing machine learning models with natural language processing techniques. This study's main objective is to create an AI sys- tem that can automatically create structured cricket scorecards from live or recorded audio commentary and use those scores to forecast the outcome of a match.

Users can submit input in the form of an audio file with live cricket commentary or a real-time voice stream recorded using a microphone. The system uses Streamlit, a robust web framework, as its frontend interface. Speech recognition is used to process the audio input and turn the spoken commentary into text, along with the names of the teams and venue information. The scorecard can then be updated continuously by appending this text to an existing dataset of match commentary.

The system uses tokenization and other NLP approaches to extract important match events, such as runs, wickets, and other noteworthy actions, once the commentary has been converted to text. A dynamic scoreboard that displays the teams' current scores, runs, and wickets taken is created by structuring the data. Additionally, a Random Forest machine learning model is used to forecast the end match score based on the real-time score data produced by the NLP system, utilizing a pre-trained T20 cricket dataset.

This study adds to the expanding body of research on AI-driven sports analytics, particularly in cricket, by investigating how machine learning models, speech recognition, and natural language processing can work together to automate the creation of cricket scorecards. A viable strategy for improving the audience experience and offering more precise insights during live cricket matches is the combination of real-time commentary analysis and predictive modeling.

## II. Review of Literature

Speech-based Emotion Recognition (SER) has emerged as a key technology in a variety of domains, including customer service optimization, mental health monitoring, and human-computer interaction. Even with these advancements, current models frequently fall short in terms of generalization beyond their training settings.We improve Whisper's speech emotion recognition capabilities and show that it outperforms HuBERT and WavLM in terms of generalization. [1]

Modern natural language processing (NLP) and speech recognition technologies enable the automated production of meeting minutes. This improves record quality by reducing errors, saving time, and facilitating rapid access to crucial information. Speech-to-writing Transcription: Helps create and make content accessible by translating spoken words into writing. [2]

In this paper, a deep learning-based framework for automatic cricket commentary is presented, utilizing TRANS-BLSTM for text production and YOLOv8 for object identification. It offers a customized cricket commentary dataset by tackling issues like variable-length inputs and a dearth of datasets. Accurate player identification and context-rich commentary, which achieved 72% accuracy and advanced sports analysis automation, are among the main accomplishments. [3]

Sports analysis and video captioning innovations are combined in automatic cricket commentary. Early techniques, such as CNNs and LSTMs, had trouble with variable-length inputs and lacked player-specific insights. In order to address these problems, this study introduces a bespoke dataset for producing precise, player-aware, and context-rich commentary from cricket footage using YOLOv8 and TRANS-BLSTM. [4]

In order to efficiently handle linguistic and contextual problems, machine translation has progressed from statistical techniques to neural structures like Seq2Seq and transform- ers. Bilingual dictionaries and cultural awareness are recent breakthroughs that improve translation quality, especially for low-resource languages. For applications like as sports commentary, real-time systems that use LSTM and transformers concentrate on striking a balance between speed and accuracy. These developments increase accessibility for a wide range of international audiences by fostering inclusion and context-aware solutions. [5]

Novel applications in sports, especially cricket, have been made possible by developments in computer vision and machine learning. To automate operations like scorecard updates, gesture recognition systems have been built using methods like logistic regression and Haar-cascade classifiers. These solutions, which offer efficiency and real-time accuracy, take the role of conventional manual procedures. In order to handle both static and dynamic motions, vision-based techniques are developing, which will make gaming more equitable and skilled. [6]

## III. METHODOLOGY

The primary goal of this study stage is to create a scoreboard by using an audio file of cricket commentary as input. The scoreboard is then used to forecast how many runs could be scored in the remaining innings. In general, our model may be divided into three sections, which are 1)Whisper AI 2)Natural language processing and techniques 3)Machine learning model. Whisper AI is used to convert the audio file of cricket commentary into the text file. NLP techniques like tokenization and lemmatization are used to break long sentences into tokens and detect the key words which are used in cricketing terminology. This helps in generating an appropriate and accurate scorecard and lastly the ML model is used to predict the final score by taking key inputs from the scorecard.

### A. Datasets

We used two distinct datasets: one is audio recordings of cricket commentary from the Cricbuzz app, and the other is a dataset from the Kaggle website that we used to create a model that predicts the final score in each innings.

URL: https://www.kaggle.com/datasets/veeralakrishna/cricsheet-a-retrosheet-for-cricket/data

### B. Data Preprocessing

Several features in the dataset used for the machine learn- ing model were eliminated utilizing feature engineering approaches because they were unnecessary for our project. Both recorded audio commentary files and live commentary can be used to provide input to the whisper model.

### C. Whisper OpenAI Model

OpenAI's Whisper AI model is a powerful automatic speech recognition (ASR) tool. Each subcomponent of its design and overall operation is in charge of a distinct aspect of the speech-to-text pipeline. [1]

The following steps can be used to decompose how the model operates:

1. Input Processing-Raw audio is transformed into a format that is appropriate for model processing by input processing .By dividing audio input into 30-second parts, the approach makes processing and analysis easier.
2. Encoder Function: The encoder part of the transformer architecture receives the log Mel Spectrogram. The encoder uses a hierarchical processing structure to extract The encoder highlights important patterns and information in the audio stream by performing feature extraction on the log Mel Spectrogram pertinent characteristics at various abstraction levels.
3. Decoder Processing: Next, the decoder component receives the encoder's output. The decoder has been trained to provide matching transcriptions or captions for the provided audio input. [2]

### D. Natural Language Processing

A branch of artificial intelligence called natural language processing (NLP) is concerned with how computers and human languages interact. It entails creating models and algorithms that process, comprehend, and produce meaningful and practical human language. Once the audio file has been converted to text by the whisper AI model, essential and relevant cricket terms that are needed to create an accurate scorecard are extracted using natural language processing (NLP) techniques. Tokenization and lemmatization are two methods used to retrieve the particular cricket phrase in its original form. Short text documents tend to be less topic-centric and noisy. Some recommended methods for working with brief text documents are make the brief text longer by using search engines.[4] The following procedures are used to use NLP approaches to extract meaningful cricket phrases from the converted text file. Step 1: Text Preprocessing Preprocess the input text to standardize and tidy it before extracting cricket terminology: Put the text in lowercase:

guarantees consistency (for example, ”BOWLER” becomes ”bowler”). Eliminate noise: Remove any numbers, punctuation, or superfluous symbols that could obstruct tokenization. Tokenize the text: Divide the input into tokens, which are words or sentences.

Lemmatization in Step 2 Lemmatization is the process of breaking words down to their most basic forms. ”bowling” → ”bowl” is one example. By using this step, you can find related terms that you might otherwise overlook because of tense or inflection discrepancies. POS Tagging in Step Three To give each token a grammatical category, use a POS tagging tool. This aids in sorting tokens according to their function within the sentence. For instance: NN and NNS nouns: ”batsman,” ”bowler,” and ”pitch.” VB and VBG verbs: ”batting,” ”bowling.” Results can be enhanced by concentrating on the specific nouns and verbs that are frequently used in cricket terminology. The cricket-related data from the processed text file is separated using the previously described NLP approaches, and a scorecard is created utilizing the pertinent and necessary data.

### *E. Prediction Model*

To train the model, we divided the data into train and test sets. We trained with 80% and tested with 20%. The ’runsx’ column, which is used for prediction, is assigned the ’y’ portion of the dataset, while the ’x’ portion is allotted the remaining dataset. We used the XGBoost machine learning technique to train the model, and the accuracy rate was 96.2%.After the model has been fully trained with the highest degree of accuracy, it is ready to forecast the output. The final score of an innings in a T20 international match is predicted  by a number of criteria, such as the batting and bowling teams, the current score, the number of overs bowled, the score from the previous five overs, and the number of wickets dropped. Our program then uses the current conditions of the match to forecast the score.

## IV. Result

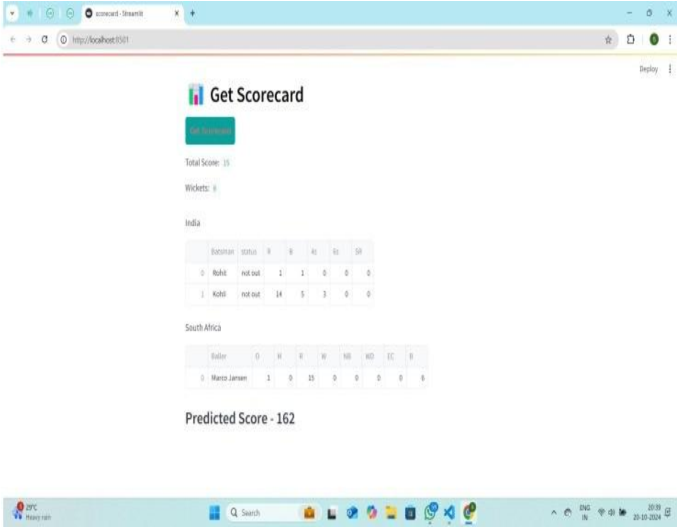
Our final product is focused on creating a cricket scorecard that includes all the information, such as the batsman’s score, bowling numbers, overs bowled, wickets fallen, and the innings’ final score. To do this, we have utilized the ’re’ library, also known as RegX (Regular Expression), which is a string of characters used to look for patterns in sentences. For the goal of natural language processing, we have employed Whisper AI.

## V. Conclusion

This study shows that it is possible to automate the creation of cricket scorecards using live or recorded commentary by utilizing machine learning, natural language processing, and artificial intelligence (AI). The system can continually update and track the progress of a cricket match in real-time by utilizing tokenization techniques for text analysis and speech recognition for audio-to-text conversion. Additionally, using the dynamically created scorecard to estimate the final match score is made possible by the use of a Random Forest machine learning model that was trained on a T20 cricket dataset. By providing insights into possible match outcomes, this predictive element provides value and enhances the overall cricket viewing experience for both fans and experts.

TABLE I
Predicted Score v/s Actual Score of various International Matches

| Batting | Bowling | Overs | Cr.Runs | Predicted Score | Actual Score |
|---|---|---|---|---|---|
| India | Afghanistan | 17 | 150 | 171 | 181 |
| Pakistan | England | 9 | 83 | 164 | 158 |
| India | Sri Lanka | 15 | 102 | 138 | 137 |
| West Indies | Australia | 18 | 189 | 204 | 207 |
| New Zealand | Pakistan | 17 | 195 | 227 | 226 |
| Afghanistan | India | 16 | 120 | 152 | 157 |
| Australia | South Africa | 15 | 165 | 186 | 191 |
| Pakistan | New Zealand | 9 | 93 | 191 | 192 |
| India | Sri Lanka | 17 | 128 | 170 | 162 |



Get Scorecard

Total Score: 15

Wickets: 0

India

| | Batsman | status | R | B | 4s | 6s | SR |
|---|---|---|---|---|---|---|---|
| 0 | Rohit | not out | 1 | 1 | 0 | 0 | 0 |
| 1 | Kohli | not out | 14 | 5 | 3 | 0 | 0 |

South Africa

| | Baller | O | M | R | W | NB | WD | EC | B |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Marco Jansen | 1 | 0 | 15 | 0 | 0 | 0 | 0 | 6 |

Predicted Score - 162

Fig. 1. Final Result

## VI. Future Work

In the future, we can employ a multilingual commentary to generate the score card, as cricket commentary is now available in other languages. People from all areas of society will be involved. By utilizing YOLOv8 and LSTM, which can recognize objects and gestures in the frame and translate them into audio or text, we can also create a video-to-text scorecard that converts pictures and objects into cricket scores.

## References

[1] Erik Goron, Lena Asai, Elias Rut, Martin Dinov, “Improving Domain Generalization in Speech Emotion Recognition with Whisper”, CASSP - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024

[2] Renu Kachhoria, Netal Daga, Himanshu Ramteke, Yash Akotkar, Samarth Ghule, ”Minutes of Meeting Generation for Online Meetings Using NLP & ML Techniques”, International Conference on Emerging Smart Computing and Informatics (ESCI) AISSMS Institute of Information Technology, Pune, India. Mar 5-7, 2024

[3] Karnati Sai Shashank, N. Praneeth Prasad, K. Spoorthy Reddy, L. Sridhara Rao, “Upload Cricket Match Video to Generate Audio Commentary by YOLOv8 and Transformer”, International Conference on Sustainable Computing and Smart Systems (ICSCSS), 2023

[4] Swarup Ranjan Behera, Parag Agrawal†, Amit Awekar, V. Vijaya Saradhi, “Mining Strengths and Weaknesses of Cricket Players Using Short Text Commentary”, 18th IEEE International Conference on Machine Learning and Applications (ICMLA), 2019.

[5] Gaurav Kamath, Udayabhaskara N, “Multilingual Transformer for Dynamic Cricket Commentary Generation”, International Conference on Computing and Data Science (ICCDS), 2024.

[6] Md. Asif Shahjalal, Zubaer Ahmad, Rushrukh Rayan, Lamia Alam, “An Approach to Automate the Scorecard in Cricket with Computer Vision and Machine Learning”, 3rd International Conference on Electrical Information and Communication Technology (EICT), 7-9 December 2017, Khulna, Bangladesh, 2017.

[7] S. R. Iyer and R. Sharda, “Prediction of athletes performance using neural networks: An application in cricket team selection,” Expert Syst. Appl., vol. 36, no. 3, pp. 5510–5522, 2009.

[8] K. Dixit and Stanford, “Deep learning using cnns for ball-by-ball outcome classification in sports,” 2016.

[9] Cricket commentary, “IND vs SL T20 match”, 2023. [Online]. Available: https://www.espncricinfo.com/series/sri-lanka-in-india- 2022-23-1348629/india-vs-sri-lanka-3rd-t20i-1348642/full-scorecard

[10] C. Li, Y. Duan, H. Wang, Z. Zhang, A. Sun, and Z. Ma, “Enhancing topic modeling for short texts with auxiliary word embeddings,” ACM Trans. Inf. Syst., vol. 36, no. 2, pp. 1101–1130, 2017.

[11] Gaurav Kamath, and Udayabhaskara N. (2024). IPL 2023 Match wise Ball by Ball Commentary [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/7918558

[12] Dr.M.Ramalingam,Mr.S.Gokul,Mr.L.S.Mythravarshini, Ms.K.S.Harine, “Efficient Player Prediction and Suggestion using Machine Learning for IPL Tournament”, International Mobile and Embedded Technology Conference (MECON), 2022

[13] Abdul Basit, Muhammad Bux Alvi, Fawwad Hassan Jaskani, MajdahAlvi, Kashif H. Memon, Rehan Ali Shah, IEEE 23rd International Multitopic Conference(INMIC), 2020

[14] Shristi Priya, Ankit Kumar Gupta, Atman Dwivedi, Aryan Prabhakar, “Analysis and Winning Prediction in T20 Cricket using Machine Learning”, Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), 2022

[15] Eeshan Mundhe, Ishan Jain, Sanskar Shah, “Live Cricket Score Prediction Web Application using Machine Learning”, International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON) Pune, India, 2021.