STARC '23 Abstract

Creating Data Repositories for sports based on video/image data

Sport:	Cricket
Leader:	Shubham Mookim(PES2UG21EC137)
Members:	• Shubham Mookim (PES2UG21EC137) • Yash Kumar (PES2UG21CS930) • Shushanth Prem Anand (PES2UG21CS518)

Project Details

Background

The lack of publicly available, curated datasets in multiple sports hinders progress in sports analytics research and development. A research area that could address this issue is the development of efficient and standardized methods for collecting and curating large volumes of sports data, with a focus on ensuring data quality, privacy, and security. Such methods could enable the creation of a centralized, open-access data repository that would facilitate algorithm development and validation, as well as the exploration of new research questions and approaches in sports analytics.

Goals, Objectives and Scope

- Develop efficient and standardized methods for collecting and curating large volumes of sports data.
- Ensure data quality, privacy, and security in the process of collecting and curating sports data.
- Create a centralized, open-access data repository for sports analytics research.
- Facilitate algorithm development and validation through access to standardized datasets.
- Foster collaboration and knowledge sharing among researchers and experts in sports analytics.
- Promote transparency and reproducibility in sports analytics research.
- Drive innovation in sports analytics by enabling exploration of new research questions and methodologies.
- Enhance performance analysis, decision-making, and strategy optimization in sports through improved data access and insights.
- Establish protocols and safeguards to protect sensitive information and comply with privacy regulations.

Applications:

1. Performance Analysis: Sports teams and coaches can utilize the curated datasets to perform in-depth analysis of player performance, team dynamics, and game strategies. This can help identify strengths, weaknesses, and patterns, leading to improved training programs, game plans, and player selections.
2. Player Development: With access to comprehensive sports data, player development programs can be enhanced. Coaches and trainers can analyze individual player performance metrics, track progress over time, and identify areas for improvement. This can contribute to the overall growth and skill development of athletes.
3. Injury Prevention and Rehabilitation: Curated datasets can assist sports science professionals in studying injury patterns, risk factors, and recovery processes. By analyzing the data, injury prevention strategies can be developed, and rehabilitation programs can be tailored based on evidence-based insights.
4. Talent Identification and Scouting: Sports analytics research using standardized datasets can help identify talented players and prospects. Data-driven analysis can provide valuable insights into player potential, performance trends, and key attributes, aiding in effective talent identification and scouting processes.

5. Game Strategy Optimization: Coaches and teams can leverage curated sports datasets to optimize game strategies. By analyzing historical data and real-time information, teams can make informed decisions regarding formations, player positions, substitutions, and in-game tactics to gain a competitive edge.
6. Fan Engagement and Media Coverage: Access to curated sports data can enhance the fan experience by providing detailed statistics, visualizations, and interactive platforms for fans to engage with the sport. Media outlets can leverage the data to provide comprehensive coverage, insights, and storytelling, enhancing the overall sports viewing experience.
7. Betting and Fantasy Sports: The availability of curated datasets can contribute to the accuracy and fairness of sports betting and fantasy sports platforms. Standardized data can enable more reliable and transparent algorithms for odds calculation, player valuation, and game predictions, providing a more engaging and satisfying experience for users.

Deliverables

1. Data Collection

To collect data, we would need to know how to gather data from various sources, including structured and unstructured data. This could involve using web scraping, APIs, or file formats such as CSV, JSON, or XML.

2. Data Cleaning and Preprocessing

To ensure the quality and accuracy of the data, we would need to be able to clean and preprocess the data. This could involve using tools such as Python, R, or SQL to manipulate and transform the data.

3. Data Storage and Management

We would need to be proficient in using databases and cloud storage solutions to store and manage the data. We may also need to know how to use version control systems such as Git to track changes to the data.

4. Data Visualization and Analysis

To analyze the data using statistical and machine learning techniques and visualize the results, we would need to be able to use tools such as Python libraries like matplotlib, seaborn, or Tableau.

5. Sports Knowledge

Having a strong understanding of the sports we are studying is essential for being able to interpret the data and develop meaningful insights. This could involve having a background in sports science, sports analytics, or sports journalism.

6. Communication

Being able to communicate our findings and insights effectively to both technical and non-technical audiences is crucial. This could involve creating reports, presentations, and visualizations that are easy to understand and interpret.

7. Collaboration

Working on a project of this nature would likely involve collaborating with other researchers, data scientists, and analysts. Therefore, being able to work effectively in a team environment is crucial.

8. Tools and Technologies

Some useful tools that we may need to use include Python, SQL, cloud storage solutions like Amazon S3, Google Cloud Storage, or Microsoft Azure, Git, Jupyter Notebooks, machine learning libraries such as scikit-learn, TensorFlow, or PyTorch, data visualization tools like matplotlib, seaborn, or Tableau, web scraping libraries such as Beautiful Soup or Scrapy, and APIs such as Twitter API, NBA API, or NFL API.

Potential Obstacles

- **1. Limited access to data:** We may face challenges in accessing high-quality and relevant data due to restrictions by data providers or sports organizations.
- **2. Data privacy and security concerns:** We may encounter obstacles related to data privacy and security, especially when dealing with sensitive information such as player performance data or injury reports.
- **3. Technical issues:** We may encounter technical issues related to data collection, cleaning, storage, and analysis that could impede our progress.

Project Approval

Suggestions

-

Approved by	Approved by:
POSITION	POSITION
DD/MM/YYYY	DD/MM/YYYY

Week #1

Meetings:

Meeting 1: Thursday(08/06/23) with Dr. Prajwala

- Presented our ideas & plan to our mentor and took the following inputs from her :
  a. To make a presentation on our action plans , framework , etc.
  b. Search for Datasets on the internet to work on .
  c. Guided us to make a Web based solution for easier access to the model we're working on.

Progress Made:

- We have had internal meetings with our group members on choosing of the datasets and have started to work towards our model.
- We have found a few reliable websites and have started to extract data from it .
- We are also researching & learning our ways to use computer vision that has to be implemented on to our project.

Next Scheduled Meeting: On 14th June With Dr. Prajwala & Sandesh sir (online)

Work to do in upcoming week :

1. Collection of data from various sources, both structured and unstructured. This will involve web scraping, APIs extractions.
2. Cleaning and preprocess the data to ensure its quality and accuracy involving tools such as Python, R & SQL.
3. Go through courses on Computer Vision & Data cleaning.

Week #2

Meetings:

Meeting 2: Wednesday(14/06/23) with Dr. Prajwala & Sandesh sir

- Mentors instructed to create a cricket data repository and visualization.
- Goal: visualize player performance based on runs scored and wickets taken.
- Team advised to check availability of cricket match videos and commentary.
- Research on Multi-Modal Systems.
- Follow-up meeting scheduled to discuss any issues encountered during data collection.

Progress Made:

- Researched and started to work with multi-modal systems .
- Imported entire match video into a cloud to work on.
- Imported the transcribed commentaries of entire cricket matches videos.

Next Scheduled Meeting: On 21st June with Dr. Prajwala (Offline)

Work to do in upcoming week :

- Set up a repository to store the data and code for the project.
- Use video analysis tools and natural language processing techniques to extract data from the collected videos and commentary.
- Clean and preprocess the data to ensure accuracy and consistency, and handle any missing or incomplete data.

Week #3

Meetings:

Meeting 3: Wednesday(21/06/2023) with Dr. Prajwala (offline)

- Our team discussed the progress made with our project with ma'am.
- Mentor instructed to breakdown our project into 3 parts :
  a. Video analysis (using computer vision & ML)
  b. Image analysis (using computer vision & ML)
  c. Text analysis (using NLP)

- Mentor has instructed & formulated our plan , we are gonna work with the video analysis part first.
- Team was advised to research & learn on Video Analysis of match clips .
- Follow-up meeting scheduled to discuss any issues encountered video analysis.

Progress Made:

- We have set up a git repository to store the data and code for the project.
- We have started to use video analysis tools and NLP techniques to extract data from the collected videos and commentary.

Next Scheduled Meeting: On 28th June with Dr. Prajwala (Online)

Work to do in upcoming week :

- To research & learn on video analytics using computer vision and ML.
- To start testing current algorithm.

July Week #3

Meetings:

Meeting : Friday(14/07/2023) with Dr. Prajwala (online)

- We discussed our current insights on our project with the mentor.
- Mentor suggested to make repositories and start working on with 2 countries.
- Mentor asked us to start working on the online repository as we're about to finish working on our model to gather data by next weekend

Progress Made:

- We have setup an AI image detection system model in our local computer to start detecting different requirements.
- We have integrated it with our current model which downloads and breaks down match video feed into frame by frame breakdown.
- We have started gathering and segregating match data into proper datasets which is our ultimate *goal.*

Next Scheduled Meeting: On 20th July with Dr. Prajwala (Online)

Work to do in upcoming week :

- Finish working on our dataset gathering model
- Continue preparing datasets for more matches (currently for 2 countries)
- Start setting up the online centralized repo for easier and wide scale access to our data.