Language and Multimodal Models in Sports: A Survey of Datasets and Applications

Haotian Xia[1], Zhengbang Yang[2], Yun Zhao[3], Yuqing Wang[4], Jingxi Li[1], Rhys Tracy[5], Zhuangdi Zhu[2], Yuan-fang Wang[5], Hanjie Chen[6]*, Weining Shen[1]*

[1]University of California, Irvine, CA, USA [2]George Mason University, VA, USA
[3]Meta Platforms, Inc., CA, USA [4]Stanford University, CA, USA
[5]University of California, Santa Barbara, CA, USA
[6] Rice University, TX, USA

{xiah6, weinings}@uci.edu, {zyang30, zzhu24}@gmu.edu, hanjie@rice.edu

Abstract

Recent integration of Natural Language Processing (NLP) and multimodal models has advanced the field of sports analytics. This survey presents a comprehensive review of the datasets and applications driving these innovations post-2020. We overviewed and categorized datasets into three primary types: language-based, multimodal, and convertible datasets. Language-based and multimodal datasets are for tasks involving text or multimodality (e.g., text, video, audio), respectively. Convertible datasets, initially single-modal (video), can be enriched with additional annotations, such as explanations of actions and video descriptions, to become multimodal, offering future potential for richer and more diverse applications. Our study highlights the contributions of these datasets to various applications, from improving fan experiences to supporting tactical analysis and medical diagnostics. We also discuss the challenges and future directions in dataset development, emphasizing the need for diverse, high-quality data to support real-time processing and personalized user experiences. This survey provides a foundational resource for researchers and practitioners aiming to leverage NLP and multimodal models in sports, offering insights into current trends and future opportunities in the field.

1 Introduction

The domain of sports, characterized by its dynamic nature and escalating popularity, stands as a testament to human endeavor and competition. As the world increasingly indulges in various sports, from global events like the FIFA World Cup to local leagues, the intersection of technology and sports has become a focal point of interest. In particular, the advancement of Natural Language Processing (NLP) has opened new avenues for enhancing user experiences (Chen et al., 2022), performance analytics (Bera et al., 2023), and operational efficiencies (Robinson, 2023).

NLP has long been instrumental in sports analytics, providing key functionalities such as game summarization (Wang et al., 2022; Huang et al., 2020; Belemkoabga et al., 2021) and commentary generation (Qi et al., 2023; Kim and Choi, 2020). These applications have contributed to enhancing fan experiences and offering analytical support to coaches and players. With the advent of large language models (LLMs) such as GPT (OpenAI, 2023), Llama2 (Touvron et al., 2023), and Gemini (Team et al., 2023), the capabilities of NLP in sports have expanded tremendously. LLMs have introduced a new level of nuance and sophistication to these applications, enriching the depth of analytics. This evolution marks a significant milestone in the use of NLP technologies in sports, pushing the boundaries of what these tools can achieve in terms of data-driven insights and user interaction. For instance, the SNIL system (Cheng et al., 2024) uses LLMs to generate sports news articles that are more insightful and closely aligned with user-provided insights, improving the overall narrative quality and user satisfaction. Furthermore, these advancements open up new possibilities for the future, where NLP can not only refine existing applications but also innovate new ways to interpret, understand, and interact with sports data. For example, achieving expert-level performance on scenario-based questions in the SportsQA (Xia et al., 2024) dataset means LLMs have sports understanding capacities close to those of sports experts. This could enhance LLM applications in areas such as AI refereeing, player mindset training, and tactics education.

Despite these advancements, there remains a notable gap in the literature: a comprehensive survey that encapsulates the breadth of datasets and applications driven by NLP and multimodal approaches within the sports domain. To the best of

*Corresponding authors.

our knowledge, while there are surveys focusing on computer vision (Zhao et al., 2023; Wu et al., 2022) in sports, primarily in movement recognition and classification, the specific contributions of NLP and its integration with multimodal data have not been thoroughly explored.

This review aims to bridge the gap by focusing on a taxonomy of datasets and applications that highlight the integration of NLP and multimodal models in sports. Particularly, our survey focuses on research and developments post-2020. Our study is organized into three principal categories: language-based datasets (§2), multimodal datasets (§3), and convertible datasets (§4)—each forming a separate section of the survey. By "convertible datasets", we refer to datasets that can be transformed into multimodal datasets through future operations such as annotation. In each section, we systematically explore different applications, initially discussing their relevance to sports and subsequently detailing the specific datasets that support these applications. It should be noted that some datasets are versatile and can be applied to multiple applications. For simplicity, we categorize them based on their primary focus in the original papers, showing in Figure 1. Additionally, an overview of the various types of datasets, including their size, source, and task type, is provided in Appendix A. We also discuss future work and challenges in §5, outlining the potential directions and obstacles in advancing the integration of NLP and multimodal models in sports.

- This is the first survey that systematically explores the integration of NLP and multimodal models in sports, setting a foundation for future academic and practical applications.

- The survey provides an extensive review of existing datasets and their applications, highlighting how advanced language models are currently utilized in the sports community.

- The survey offers a strategic roadmap for the future, suggesting how NLP and multimodal datasets can be effectively applied to sports applications to benefit the community.

By delving into these areas, this survey provides a vital resource for the sports community to harness the power of NLP and multimodal models in enhancing the sporting experience.

2 Language-Based Datasets

Language-based datasets are crucial in the evolving landscape of sports analytics and applications. By harnessing the power of NLP, these datasets enable deeper insights and more effective solutions across various domains, from medical applications and educational tools to enhancing fan engagement and understanding sports dynamics. This section introduces these datasets and illustrates their impact through specific applications.

2.1 Game and Player Performance Prediction and Analysis

Accurately predicting game outcomes and Analysis player performance is critical for teams, players, and coaches as it informs strategic planning, enhances decision-making, and optimizes performance on the field.

The dataset by Oved et al. (2020) includes 1,337 pre-game interviews and corresponding performance metrics for key NBA players over 14 seasons, totaling 5,226 interview-metric pairs. This dataset aims to predict deviations from mean in-game actions, utilizing linguistic signals from interviews to forecast strategic choices, player behavior, and risk-related decisions. Velichkov et al. (2019) presents a dataset comprising 50 articles of pre-match interviews with athletes from individual sports like boxing, MMA, and tennis. This dataset also includes structured data such as sports rankings, ages, and previous match results, providing a comprehensive basis for predicting match outcomes based on the athletes' interviews and historical performance data. Substitution in soccer is essential for winning the game. Bera et al.'s (2023) dataset combines past performance metrics and text converted from audio recordings of conversations between players and coaches for sentiment analysis to enhance substitution decisions in soccer.

2.2 Hate Speech Detection

In sports, where fans and athletes engage extensively online, hate speech can detract from the community experience and affect mental well-being. Effective detection and management of hate speech are essential to maintaining a supportive environment for all participants. This ensures that digital interactions around sports events remain respectful and inclusive, reflecting the true spirit of sportsmanship that underpins athletic competition.

The dataset provided by Vujičić Stanković and

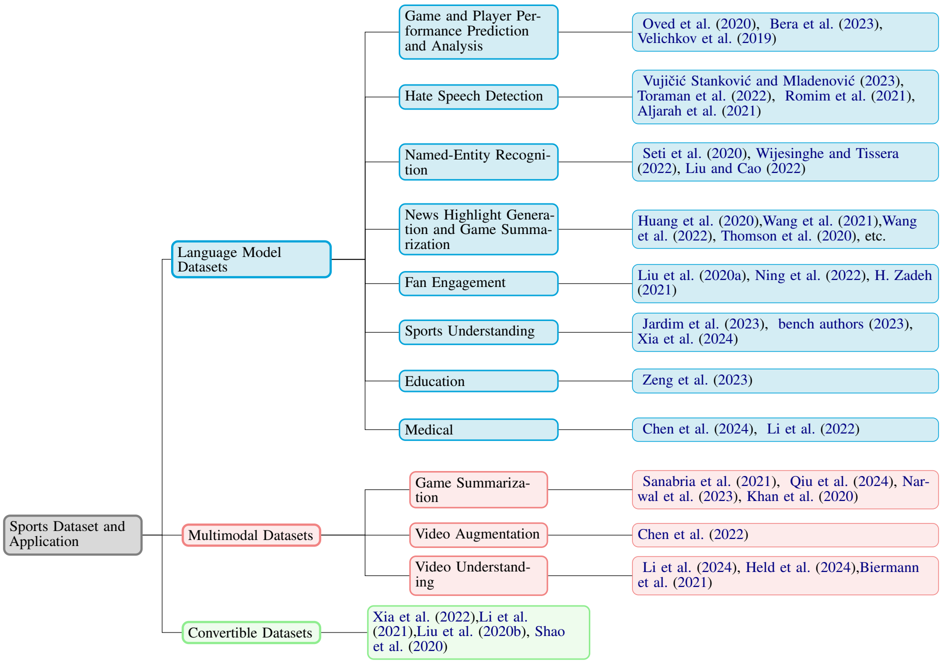
Sports Dataset and Application
Language Model Datasets
Game and Player Performance Prediction and Analysis
Oved et al. (2020), Bera et al. (2023), Velichkov et al. (2019)
Hate Speech Detection
Vujičić Stanković and Mladenović (2023), Toraman et al. (2022), Romim et al. (2021), Aljarah et al. (2021)
Named-Entity Recognition
Seti et al. (2020), Wijesinghe and Tissera (2022), Liu and Cao (2022)
News Highlight Generation and Game Summarization
Huang et al. (2020),Wang et al. (2021),Wang et al. (2022), Thomson et al. (2020), etc.
Fan Engagement
Liu et al. (2020a), Ning et al. (2022), H. Zadeh (2021)
Sports Understanding
Jardim et al. (2023), bench authors (2023), Xia et al. (2024)
Education
Zeng et al. (2023)
Medical
Chen et al. (2024), Li et al. (2022)
Multimodal Datasets
Game Summarization
Sanabria et al. (2021), Qiu et al. (2024), Narwal et al. (2023), Khan et al. (2020)
Video Augmentation
Chen et al. (2022)
Video Understanding
Li et al. (2024), Held et al. (2024),Biermann et al. (2021)
Convertible Datasets
Xia et al. (2022),Li et al. (2021),Liu et al. (2020b), Shao et al. (2020)

Figure 1: Taxonomy of Research on Sports Datasets and Corresponding Applications

Mladenović (2023) is designed to identify hate speech in sports-related communications on social media. The dataset includes data from multiple sources, focusing on comments published on popular entertainment and sports YouTube channels as well as sports news articles on Serbian news portals. Toraman et al. (2022) propose a dataset comprising 100,000 tweets across multiple domains, including sports, to test cross-domain model efficiency. Romim et al. (2021) and Aljarah et al. (2021) contribute by focusing on Bengali and Arabic languages, respectively, expanding the scope to include sports-related content within larger multilingual and multicultural datasets, with 30,000 user comments and 3,696 tweets analyzed, respectively.

2.3 Named-Entity Recognition

Named-Entity Recognition (NER) in sports efficiently extracts and categorizes key details like player names, team names, and event specifics from sports-related texts. This foundational technology supports complex systems beyond summarization, enhancing data analysis, fan engagement, and personalized content delivery by providing structured and accessible data.

The dataset (Seti et al., 2020) is crafted from major Chinese sports news websites and contains sports texts used to test a NER model. It features different types of entities, such as sporting event names, stadium names, team names, player names, and results. Wijesinghe and Tissera's (2022) dataset is specifically built for Sinhala language sports-related content. It includes annotated data from 2,000 Sinhala e-news articles, focusing on named entities relevant to the sports domain such as tournament names, school names, and ground locations. Liu and Cao (2022) propose a dataset that includes detailed annotations across various entity types relevant to the sports domain. Specifically, the dataset features 12,802 sentences with nine different entity types, such as athletes, coaches, venues and sports items, which are crucial for enriching the content's contextual understanding and specificity.

2.4 News Highlight Generation and Game Summarization

In the sports media field, News Highlight Generation and Game Summarization serve critical roles by distilling extensive match details into digestible and engaging content. While both processes aim to encapsulate key moments, they do so with different focuses and objectives. News Highlight Generation captures the most impactful or emotionally charged events to enhance fan engagement,emphasizing dramatic moments, key plays, and exciting highlights. Conversely, Game Summarization provides a comprehensive and concise account of the game, emphasizing pivotal plays, outcomes, and statistics to give a complete and structured overview of the match. Huang et al. (2020) introduce the SportsSum dataset, which consists of 5,428 pairs of live commentaries and corresponding news articles derived from major soccer leagues. This dataset is specifically designed to address the challenges of summarizing sports games, where the narrative style of the source (commentaries) differs from that of the target summaries (news articles). Subsequent iteration, SportsSum2.0 (Wang et al., 2021), refined SportsSum further by cleaning noisy data. Building on SportsSum2.0, the K-SportsSum (Wang et al., 2022) dataset not only increased the dataset to 7,854 commentary-news pairs but also enriched it with comprehensive background knowledge of teams and players, showcasing a pivotal development in bridging the knowledge gap essential for generating more informative sports news.

Thomson et al.'s (2020) dataset is structured to facilitate better Natural Language Generation for sports by providing a comprehensive, multi-dimensional dataset of basketball statistics. It includes detailed game data from the 2014 to 2018 NBA seasons, organized in a way that allows for complex queries and supports a richer entity relationship graph, aimed at enhancing data-driven sports content creation. Sarfati et al. (2023) propose a dataset containing 143 annotated narratives from selected Premier League games, aimed at fine-tuning a Natural Language Inference model to ensure the factual accuracy of the news narratives. The dataset in the study (Cheng et al., 2024) uses a curated dataset of 120 sports news articles to demonstrate the application of large language models in sports journalism.

2.5 Fan Engagement

Fan engagement in sports is increasingly driven by data analytics, with organizations leveraging datasets to understand and enhance the fan experience. This subsection explores various datasets that can be used by language models to quantify and improve fan engagement using text analytics and sentiment analysis.

Liu et al. (2020a) introduce the LiveQA dataset, derived from interactive content in NBA play-by-play live broadcasts, encompassing 117k multiple-choice questions from 1,670 games. This dataset captures dynamic fan interactions during live events, illustrating how engaging real-time content can boost fan participation and enhance the viewing experience. Ning et al.'s (2022) dataset consists of 821 comments totaling 1,205 lines of text collected through social media surveys from basketball season ticket holders. This dataset was used to build models for text categorization and polarity assessments via sentiment analysis, providing insights into fan sentiments that can guide engagement strategies, H. Zadeh (2021) leverage a dataset comprising 391,092 entries from Zhihu to analyze sports-related queries and discussions. This dataset facilitates a comprehensive understanding of the topics that dominate fan interactions and the types of information most sought after by different user groups within the community. This insight is essential for tailoring engagement and content strategies to meet the specific needs of sports enthusiasts effectively.

2.6 Sports Understanding

Sports Understanding is pivotal for the development of AI applications, such as AI referees, where the sophistication of model comprehension 2impacts their effectiveness and reliability. This subsection explores QA datasets that are instrumental in enhancing models' capabilities to interpret complex sports scenarios, rules, and statistics. These datasets train models to deliver a higher quality of content and interaction within the sports community, supporting both sports enthusiasts and professionals by providing deeper insights and a more intuitive understanding of sports through advanced technology. Such capabilities are essential for developing systems that can effectively analyze and respond to dynamic sports environments.

QASports (Jardim et al., 2023) is a context extractive question-answering dataset focusing on soccer, American football, and basketball. It primarily involves answering questions based on specifically provided paragraphs focused on fact-based questions. The BIG-bench sports understanding task (bench authors, 2023), is a subtask within the broader BIG-bench initiative designed to evaluate the sports-specific comprehension of large language models. This task presents 986 multiple-choice questions where models must discern plausible from implausible sports-related statements, requiring knowledge of athletes' names and actions typical in various sports. It challenges models to use higher-level reasoning to understand the nuances of sports contexts rather than merely recalling associations.

The more challenging dataset, SportQA (Xia et al., 2024), is specially designed to test the sports understanding of large language models (LLMs) across three levels of difficulty, from basic historical facts and sports strategies and rules to advanced scenario-based reasoning. It contains 70,592 multiple-choice questions covering 35 different sports that challenge models to demonstrate deep comprehension of sports-related scenarios and facts. This dataset addresses the gap in sports-specific benchmarks for evaluating the nuanced understanding of sports in LLMs.

2.7 Medical Application

Medical applications in sports, facilitated by advancements in NLP, are crucial for analyzing clinical texts to support injury diagnosis and treatment. These technologies are especially valuable in sports medicine for processing and interpreting radiological and surgical data, aiding in the timely and accurate medical care of athletes.

Knee Meniscal Tears Dataset (Li et al., 2022) involves radiology and arthroscopy reports focused on knee meniscal tears, relevant for sports due to the frequency of knee injuries among athletes. The dataset includes 3,593 knee MRI reports from a single institution, providing a substantial basis for developing NLP models that automate the extraction and correlation of medical findings.

Managing osteoarthritis is important in sports, as the condition frequently results from sports-related injuries (Bullock et al., 2020) and cans impact athletic performance and career longevity. The Osteoarthritis dataset (Chen et al., 2024) consists of data from 80 real-world patients diagnosed with osteoarthritis, collected between April and October 2023. It includes comprehensive details like age, BMI, symptoms, and radiographic findings, derived from six well-established clinical guidelines. This dataset aims to evaluate the capability of large language models to generate personalized treatment plans based on evidence-based medicine.

2.8 Educational Application

Educational applications in sports increasingly utilize NLP to enhance teaching methodologies and student engagement. This involves analyzing educational content and interactions to provide personalized learning experiences and assess teaching effectiveness. Zeng et al. (2023) explores the effectiveness of digital teaching methods in sports education by utilizing a hybrid intelligent system that combines text and visual features for quality evaluation. The datasets used in this study are essential for assessing and enhancing the interactivity and educational value of sports training videos. Specifically, the Charades Dataset, consisting over 10,000 videos with 157 action types, and the Net-caption Dataset, with over 20,000 videos and 100,000 video segment-sentence pairs, are instrumental in developing models assessing digital sports education quality.

3 Multimodal Datasets

Multimodal datasets integrate multiple types of data, such as text, video, and audio, to provide a richer and more comprehensive understanding of sports events. These datasets are crucial for developing advanced AI models that can interpret and analyze complex scenarios, offering deeper insights and more engaging experiences for both fans and professionals. In this section, we explore various applications of multimodal datasets in sports, including game summarization, video understanding, and video augmentation. Each subsection highlights specific datasets and their contributions to enhancing sports analytics and user experiences.

3.1 Game Summarization

Game Summarization leverages multimodal datasets to deliver concise and engaging sports content. These datasets encompass video, audio and textual data, providing a richer context for summarization technologies. These datasets enable the creation of video summaries that retain the original game's visual and auditory essence, enhancing the viewer's experience by highlighting crucial moments. The dataset (Sanabria et al., 2021) consists of event data from 70 soccer matches across two European leagues, including a diverse range of event types like passes, shots, and fouls. Each match generates approximately 1,700 event instances which are used to train the summarization model. This comprehensive dataset allows for a nuanced analysis of game dynamics, aiming to use Multimodal Models to automate the generation of match highlights and summaries effectively.

The MSMO dataset (Qiu et al., 2024) features 5,100 videos across various themes, with sports being one of the prominent categories. It provides both video and textual summaries, which are valuable for developing multimodal summarization techniques. This aspect makes it suited for applications that require combining visual content with descriptive text to enhance the viewer's understanding and engagement with sports events. Narwal et al. (2023) propose two multimodal datasets tailored for cricket video summarization: the DPCS (Delivery Play Cricket Sport) image dataset and the EXINP (Excited Interval Normal Play) Cricket audio dataset. The DPCS dataset contains 8,646 high-definition images divided into two classes, Delivery and Play, used for segmenting cricket videos. The EXINP dataset comprises 868 audio segments, categorized into Excited, Interval, and Normal Play, to identify key events. Together, these datasets utilize audio-visual cues for dynamic cricket video summarization.

Khan et al.'s (2020) include 104 broadcast sports videos, totaling over 230 hours of content. This dataset encompasses multiple sports types and includes additional elements like commercials, providing a comprehensive base for testing summarization models. It integrates audio and visual cues—such as crowd reactions and score changes—to identify key moments for video summarization, emphasizing its utility in creating enriched sports highlights that effectively capture the essence of events.

3.2 Video Understanding

Video understanding in sports utilizes multimodal models to interpret and analyze complex scenarios beyond the capabilities of traditional classification models. Unlike predefined task-based models, multimodal models can perform complex and open-ended tasks, such as detecting and describing whether a player has broken a rule, while CV classification models are more suited for identifying specific and well-defined movements performed by a player.

The Sports-QA dataset (Li et al., 2024) comprises approximately 94,000 question-answer pairs across four types of questions derived from sports videos and professional action labels sourced from the MultiSports and FineGym datasets. Descriptive questions involve simple queries about the actions and events in the video. Temporal questions focus on the relationships among actions over time. Causal questions uncover the reasons or processes behind actions by exploring causal relationships. Counterfactual questions set hypothetical conditions not occurring in the video and query the expected outcomes based on these conditions. The dataset's complexity and diversity make it highly adaptable to multimodal settings, providing a robust foundation for testing the model's capabilities.

The X-VARS dataset (Held et al., 2024) enhances the evaluation of multimodal models with over 10,000 video clips and more than 22,000 video-question-answer triplets sourced from football games Annotated by over 70 experienced football referees. It includes complex, high-level questions that require identification and further explanation. Another contribution is the EIGD dataset(Biermann et al., 2021). This dataset includes 125 minutes of handball and 125 minutes of soccer video. The handball videos are provided by the Deutsche Handball Liga and Kinexon, while the soccer videos are from publicly available broadcast recordings of the FIFA World Cup. The dataset features questions about low-level actions (free kicks, throw-ins) and high-level actions (passes, shots), enriching the training and evaluation of multimodal models in understanding various granularity levels in sports events.

3.3 Video Augmentation

Video augmentation in sports broadcasting harnesses multimodal datasets to enhance the viewer experience by integrating real-time data visualizations into live broadcasts. This not only enriches the visual presentation but also deepens fans' understanding of the game by providing insightful, data-driven analyses directly within the video feed. Chen et al. (2022) presents a dataset comprising 155 sports video clips paired with commentary text. This multimodal dataset is used to develop a system that augments sports videos by dynamically integrating visualized data from the commentaries into the videos. It serves as a foundation for developing practical applications of linking textual and visual data to enhance multimedia content.

4 Convertible Datasets

By "convertible datasets," we refer to datasets that are not yet multimodal but can be transformed into multimodal datasets through future operations such as annotation. We focus on the potential conversion to multimodal datasets rather than solely to language-based datasets because multimodal capabilities enable richer and more diverse applications in sports, enhancing the analysis and understanding of complex interactions within the data. The introduction of convertible datasets is particularly crucial given the relative scarcity of multimodal datasets compared to language-based datasets, underscoring the need to expand the availability and versatility of multimodal resources.

The VREN dataset (Xia et al., 2022) introduces a generalized language for describing volleyball rallies through its dataset variable settings, meeting professional tactical statistics requirements. However, not all rallies are annotated. By annotating each rally and verifying the accuracy, this dataset can become multimodal. This allows large language models to generate similar descriptions, aiding both fans in understanding matches and professionals in real-time statistical analysis, while also enhancing model understanding of volleyball.

The MultiSports dataset (Li et al., 2021) includes annotated action instances across basketball, volleyball, football, and aerobic gymnastics. The FineGym dataset (Shao et al., 2020)provides hierarchical annotations for gymnastics, structured into events, sets, and elements, with both semantic and temporal annotations. The FSD-10 dataset (Liu et al., 2020b) contains video clips from figure skating competitions, annotated with action types, base values, grades of execution, and other metadata. By adding detailed textual descriptions, integrating sensor data, and including audio commentary, these datasets can become multimodal. We focus on these three datasets as representative examples to illustrate the potential of converting sports datasets into multimodal resources through annotation. Transforming these datasets will enable them to support advanced AI models by providing deeper contextual information, improving the granularity of real-time analysis, and facilitating the development of applications that help both fans and professionals comprehend complex sports

5 Future Work and Challenges

In this section, we delve into the future directions and challenges associated with NLP and multimodal datasets in sports. As the first survey to systematically explore this integration, we emphasize the diverse applications of these technologies and the corresponding dataset requirements. While sections 2 and 3 categorized datasets based on their type and primary use, here we adopt a different approach to highlight the specific needs and potential impact on different backgrounds of people within the sports community. Our discussion is categorized into three main application areas: applications for enthusiasts, applications for professionals, and medical or rehabilitation applications. Each category not only highlights the current contributions but also outlines the challenges and future work needed to enhance these applications. Among all the challenges, data quality stands out as a critical factor that impacts the effectiveness of these applications.

5.1 Applications for Enthusiasts

Applications for sports enthusiasts primarily focus on enhancing fan engagement and understanding of the game. Key applications include highlight generation, news generation, and rule explanation augmentation. These applications are essential for maintaining and growing the fan base by providing richer, more engaging experiences.

5.1.1 Highlight Generation and News Generation

NLP and multimodal datasets have improved the generation of game highlights and sports news, making them more engaging and informative. Automated systems can now create concise and captivating highlights by analyzing game footage and generating corresponding textual summaries. These applications keep fans updated and involved, making sports more accessible and enjoyable.

Challenges: *Dataset Diversity and Quality:* Ensuring datasets cover a wide range of sports and include diverse and high-quality video and textual data is crucial. Current datasets lack the breadth required to generalize across different sports and events. *Real-Time Processing:* The need for real-time highlight generation demands efficient processing capabilities and prompt data availability, which poses a significant challenge for existing datasets. *Fan Engagement Insights:* Understanding and predicting fan preferences through data-driven insights is essential. This requires datasets that not only capture game events but also fan reactions and interactions.

5.1.2 Rule Explanation and Augmentation

Enhancing fan understanding through applications that provide detailed rule explanations and augmented game insights in real-time can elevate the viewing experience. These applications make complex rules accessible to viewers, broadening the audience base by augmenting broadcasts with informative overlays and contextual details.

Challenges: *Integration of Multimodal Data:* Combining video, audio, and textual data seamlessly to provide coherent and contextually accurate rule explanations requires advanced multimodal datasets. *Personalization:* Personalizing content based on individual viewer's knowledge level and preferences necessitates datasets that can support adaptive learning and recommendation systems.

5.2 Applications for Professionals

Applications designed for athletes, coaches, and sports analysts focus on real-time tactical analysis and performance monitoring. These applications are crucial for enhancing the competitive edge and performance of athletes and teams. Compared to applications for enthusiasts, professionals require more detailed, accurate, and high-quality data, as well as different focuses. For example, while both may benefit from highlight generation, professionals need highlights that focus on specific tactics and player performance under certain conditions, offering a more nuanced and detailed perspective.

5.2.1 Tactical Analysis and Performance Monitoring

NLP and multimodal datasets enable detailed analysis of player performance and team tactics, assisting coaches in making decisions during games. These applications include real-time data analytics, motion analysis, and tactical simulations. They are vital for optimizing performance and strategy.

Challenges: *Real-Time Data Collection:* Accurate and timely data collection during games is critical. Current datasets need to improve in capturing real-time data efficiently. *High Precision and Reliability:* Applications for professionals demand high precision and reliability, which requires datasets with extensive and accurate annotations. *Integration of Various Data Types:* Integrating biometric, video, and textual data to provide comprehensive insights is complex and requires sophisticated datasets.

5.2.2 Training and Simulation Tools

Advanced training tools that simulate game scenarios and provide feedback on player performance can enhance training efficiency. These tools rely heavily on accurate datasets that reflect real-game conditions.

Challenges: *Realism in Simulations:* Ensuring simulations are realistic and reflect actual game conditions requires highly detailed and contextually rich datasets. *Adaptive Learning:* Training tools need to adapt to the evolving skills of athletes, which requires datasets that can support continuous learning and adaptation.

5.3 Medical and Rehabilitation Applications

Applications in the medical domain focus on injury diagnosis, treatment planning, and psychological support for athletes. These applications ensure the health and longevity of athletes' careers.

5.3.1 Injury Diagnosis and Treatment

NLP and multimodal datasets play a crucial role in analyzing medical reports, imaging data, and clinical records to support injury diagnosis and treatment planning. Accurate diagnosis and effective treatment plans are essential for the quick recovery and sustained performance of athletes.

Challenges: *Comprehensive Medical Datasets:* Creating comprehensive datasets that cover various types of sports injuries and include detailed medical histories, imaging data, and treatment outcomes is essential. *Data Privacy and Security:* Ensuring the privacy and security of medical data while making it accessible for research and application development is a significant challenge. *Expert Validation:* Collecting and validating medical data requires input from medical professionals to ensure accuracy and relevance, adding a layer of complexity to dataset creation.

5.3.2 Psychological Support and Rehabilitation

Providing psychological support and monitoring rehabilitation progress through NLP-driven applications can aid athletes in their recovery process. These applications benefit from datasets that include psychological assessments, treatment plans, and recovery progress.

Challenges: *Holistic Data Integration:* Integrating psychological, physiological, and performance data to provide comprehensive support requires multifaceted datasets. *Personalization and Adaptation:* Datasets need to support personalized treatment and adaptation based on individual recovery trajectories.

6 Conclusion

In this paper, we systematically reviewed applications and various datasets in integrating NLP and multimodal models within the sports domain. We categorized datasets into language-based, multimodal, and convertible types, highlighting their applications and associated challenges. We discussed future work needed to enhance applications for enthusiasts, professionals, and medical or rehabilitation purposes. By addressing these challenges, we aim to inspire further advancements and practical applications that meet the evolving needs of the sports community.

Limitations

While this survey provides a comprehensive review of the integration of NLP and multimodal models in sports, several limitations should be acknowledged.

Firstly, the categorization of datasets into specific applications is based on the primary focus of their respective papers. However, many datasets have the potential to be applied across multiple domains. We mainly categorize them based on their main focus in the original papers, aiming for clarity and precision, though this may occasionally reduce the comprehensiveness of our classification. Secondly, the survey focuses on datasets and developments post-2020, potentially excluding valuable earlier contributions that could still be relevant. This temporal limitation is intended to highlight recent advancements but may inadvertently overlook foundational work that laid the groundwork for current innovations. Lastly, while we have highlighted key datasets and applications, the rapid evolution of both sports and technology means that new datasets and models are continuously being developed. Therefore, this survey represents a snapshot in time and should be updated regularly to reflect ongoing progress in the field.

References

Ibrahim Aljarah, Maria Habib, Neveen Hijazi, Hossam Faris, Raneem Qaddoura, Bassam Hammo, Mohammad Abushariah, and Mohammad Alfawareh. 2021. Intelligent detection of hate speech in arabic social network: A machine learning approach. *Journal of Information Science*, 47(4):483–501.

David Stéphane Belemkoabga, Aurélien Bossard, Abdallah Essa, Christophe Rodrigues, and Kévin Sylla. 2021. Neural network-based generation of sport summaries: A preliminary study. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 147–154, Held Online. INCOMA Ltd.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Aditya Bera, Savio Joshi, and Akhil M Nair. 2023. Football player substitution analysis using nlp and survival analysis. In *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 753–757. IEEE.

Henrik Biermann, Jonas Theiner, Manuel Bassek, Dominik Raabe, Daniel Memmert, and Ralph Ewerth. 2021. A unified taxonomy and multimodal dataset for events in invasion games. In *Proceedings of the 4th International Workshop on Multimedia Content Analysis in Sports*, pages 1–10.

Garrett S Bullock, Gary S Collins, Nick Peirce, Nigel K Arden, and Stephanie R Filbay. 2020. Playing sport injured is associated with osteoarthritis, joint pain and worse health-related quality of life: a cross-sectional study. *BMC musculoskeletal disorders*, 21:1–11.

Xi Chen, MingKe You, Li Wang, WeiZhi Liu, Yu Fu, Jie Xu, Shaoting Zhang, Gang Chen, and Jian Li. 2024. Evaluating and enhancing large language models performance in domain-specific medicine: Osteoarthritis management with docoa. *arXiv preprint arXiv:2401.12998*.

Zhutian Chen, Qisen Yang, Xiao Xie, Johanna Beyer, Haijun Xia, Yingcai Wu, and Hanspeter Pfister. 2022. Sporthesia: Augmenting sports videos using natural language. *IEEE transactions on visualization and computer graphics*, 29(1):918–928.

Liqi Cheng, Dazhen Deng, Xiao Xie, Rihong Qiu, Mingliang Xu, and Yingcai Wu. 2024. Snil: Generating sports news from insights with large language models. *IEEE Transactions on Visualization and Computer Graphics*.

Amir H. Zadeh. 2021. Quantifying fan engagement in sports using text analytics. *Journal of Data, Information and Management*, 3(3):197–208.

Jan Held, Hani Itani, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. 2024. X-vars: Introducing explainability in football refereeing with multi-modal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3267–3279.

Kuan-Hao Huang, Chen Li, and Kai-Wei Chang. 2020. Generating sports news from live commentary: A Chinese dataset for sports game summarization. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 609–615, Suzhou, China. Association for Computational Linguistics.

Pedro Calciolari Jardim, Leonardo Mauro Pereira Moraes, and Cristina Dutra Aguiar. 2023. Qasports: A question answering dataset about sports. In *Anais do V Dataset Showcase Workshop*, pages 1–12. SBC.

Abdullah Aman Khan, Jie Shao, Waqar Ali, and Saifullah Tumrani. 2020. Content-aware summarization of broadcast sports videos: an audio–visual feature extraction approach. *Neural Processing Letters*, 52(3):1945–1968.

Byeong Jo Kim and Yong Suk Choi. 2020. Automatic baseball commentary generation using deep learning. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 1056–1065.

Haopeng Li, Andong Deng, Qiuhong Ke, Jun Liu, Hossein Rahmani, Yulan Guo, Bernt Schiele, and Chen Chen. 2024. Sports-qa: A large-scale video question answering benchmark for complex and professional sports. *arXiv preprint arXiv:2401.01505*.

Matthew D Li, Francis Deng, Ken Chang, Jayashree Kalpathy-Cramer, and Ambrose J Huang. 2022. Automated radiology-arthroscopy correlation of knee meniscal tears using natural language processing algorithms. *Academic radiology*, 29(4):479–487.

Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. 2021. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13536–13545.

Pingshan Liu and Yuan Cao. 2022. A named entity recognition method for chinese winter sports news based on roberta-wwm. In *2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, pages 785–790. IEEE.

Qianying Liu, Sicong Jiang, Yizhong Wang, and Sujian Li. 2020a. LiveQA: A question answering dataset over sports live. In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 1057–1067, Haikou, China. Chinese Information Processing Society of China.

Shenlan Liu, Xiang Liu, Gao Huang, Lin Feng, Lianyu Hu, Dong Jiang, Aibin Zhang, Yang Liu, and Hong Qiao. 2020b. Fsd-10: a dataset for competitive sports content analysis. *arXiv preprint arXiv:2002.03312*.

Pulkit Narwal, Neelam Duhan, and Komal Kumar Bhatia. 2023. A novel multi-modal neural network approach for dynamic and generic sports video summarization. *Engineering Applications of Artificial Intelligence*, 126:106964.

Chuanlin Ning, Jian Xu, Hao Gao, Xi Yang, and Tianyi Wang. 2022. Sports information needs in chinese online q&a community: topic mining based on bert. *Applied Sciences*, 12(9):4784.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Nadav Oved, Amir Feder, and Roi Reichart. 2020. Predicting in-game actions from interviews of nba players. *Computational Linguistics*, 46(3):667–712.

Ji Qi, Jifan Yu, Teng Tu, Kunyu Gao, Yifan Xu, Xinyu Guan, Xiaozhi Wang, Bin Xu, Lei Hou, Juanzi Li, et al. 2023. Goal: A challenging knowledge-grounded video captioning benchmark for real-time soccer commentary generation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5391–5395.

Jielin Qiu, Jiacheng Zhu, William Han, Aditesh Kumar, Karthik Mittal, Claire Jin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Ding Zhao, et al. 2024. Mmsum: A dataset for multimodal summarization and thumbnail generation of videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21909–21921.

Drew Robinson. 2023. Evaluating the potential of ai in sports consulting: Investigating chatgpt-4's ability to consult an mlb team.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2021. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020*, pages 457–468. Springer.

Melissa Sanabria, Frédéric Precioso, and Thomas Menguy. 2021. Hierarchical multimodal attention for deep video summarization. In *2020 25th International conference on pattern recognition (ICPR)*, pages 7977–7984. IEEE.

Noah Sarfati, Ido Yerushalmy, Michael Chertok, and Yosi Keller. 2023. Generating factually consistent sport highlights narrations. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, MMSports '23, page 15–22, New York, NY, USA. Association for Computing Machinery.

Xieraili Seti, Aishan Wumaier, Turgen Yibulayin, Diliyaer Paerhati, Lulu Wang, and Alimu Saimaiti. 2020. Named-entity recognition in sports field based on a character-level graph convolutional network. *Information*, 11(1):30.

Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. 2020. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2616–2625.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Craig Thomson, Ehud Reiter, and Somayajulu Sripada. 2020. SportSett:basketball - a robust and maintainable data-set for natural language generation. In *Proceedings of the Workshop on Intelligent Information Processing and Natural Language Generation*, pages 32–40, Santiago de Compostela, Spain. Association for Computational Lingustics.

Cagri Toraman, Furkan Şahinuç, and Eyup Halit Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. *arXiv preprint arXiv:2203.01111*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Boris Velichkov, Ivan Koychev, and Svetla Boytcheva. 2019. Deep learning contextual models for prediction of sport event outcome from sportsman's interviews. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1240–1246, Varna, Bulgaria. INCOMA Ltd.

Staša Vujičić Stanković and Miljana Mladenović. 2023. An approach to automatic classification of hate speech in sports domain on social media. *Journal of Big Data*, 10(1):109.

Jiaan Wang, Zhixu Li, Qiang Yang, Jianfeng Qu, Zhigang Chen, Qingsheng Liu, and Guoping Hu. 2021. Sportssum2. 0: Generating high-quality sports news from live text commentary. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3463–3467.

Jiaan Wang, Zhixu Li, Tingyi Zhang, Duo Zheng, Jianfeng Qu, An Liu, Lei Zhao, and Zhigang Chen. 2022. Knowledge enhanced sports game summarization. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1045–1053.

WMSK Wijesinghe and Muditha Tissera. 2022. Sinhala named entity recognition model: Domain-specific classes in sports. In *2022 4th International Conference on Advancements in Computing (ICAC)*, pages 138–143. IEEE.

Fei Wu, Qingzhong Wang, Jiang Bian, Ning Ding, Feixiang Lu, Jun Cheng, Dejing Dou, and Haoyi Xiong. 2022. A survey on video action recognition in sports: Datasets, methods and applications. *IEEE Transactions on Multimedia*.

Haotian Xia, Rhys Tracy, Yun Zhao, Erwan Fraisse, Yuan-Fang Wang, and Linda Petzold. 2022. Vren: volleyball rally dataset with expression notation language. In *2022 IEEE International Conference on Knowledge Graph (ICKG)*, pages 337–346. IEEE.

Haotian Xia, Zhengbang Yang, Yuqing Wang, Rhys Tracy, Yun Zhao, Dongdong Huang, Zezhi Chen, Yan Zhu, Yuan-fang Wang, and Weining Shen. 2024. Sportqa: A benchmark for sports understanding in large language models. *arXiv preprint arXiv:2402.15862*.

Boyi Zeng, Jun Zhao, and Shantian Wen. 2023. A textual and visual features-jointly driven hybrid intelligent system for digital physical education teaching quality evaluation. *Mathematical Biosciences and Engineering*, 20(8):13581–13601.

Zhonghan Zhao, Wenhao Chai, Shengyu Hao, Wenhao Hu, Guanhong Wang, Shidong Cao, Mingli Song, Jenq-Neng Hwang, and Gaoang Wang. 2023. A survey of deep learning in sports applications: Perception, comprehension, and decision. *arXiv preprint arXiv:2307.03353*.

A Appendix

Table 1 shows the comprehensive overview of language-based, multimodal, and convertible datasets for sports applications.

Dataset	Size	Data Source	Task Type
Language-Based Dataset			
Li et al. (2022)	3,593	MRI reports	Detection
Chen et al. (2024)	80	Clinical guidelines	Open-ended-QA
Zeng et al. (2023)	110,000	Combination of different datasets	Feature Extraction & Evaluation+Description Generation
Vujičić Stanković and Mladenović (2023)	180,785	YouTube and News Portals comments	Classification(yes/no)
Toraman et al. (2022)	200,000	Tweets	Classification
Romim et al. (2021)	30,000	YouTube and Facebook comments	Classification
Aljarah et al. (2021)	3,696	Arabic tweets	Classification
Seti et al. (2020)	97,211	Combination of existing datasets	Named-entity Recognition
Wijesinghe and Tissera (2022)	99,972	Sinhala sport news	Named-entity Recognition
Liu and Cao (2022)	N/A	News portals	Named-entity Recognition
Huang et al. (2020)	5,428	Sina Sports Live	Summarization Generation
Wang et al. (2021)	5,402	Chinese sports websites	News Generation
Wang et al. (2022)	7,854	Comments from Sina Sports Live	Summarization Generation
Thomson et al. (2020)	N/A	Rotowire, basketball-reference and Wikipedia	Summarization Generation
Sarfati et al. (2023)	N/A	Narrative Generation	
Cheng et al. (2024)	N/A	NBA Games	Narrative Generation
Liu et al. (2020a)	117,000	Hupu Section of NBA Games	Multiple-Choice-QA
Ning et al. (2022)	391,092	Zhihu Section of Sports	Used to Study the Sport Community
H. Zadeh (2021)	821	Survey of basketball Season Ticket Holders	Text Mining and Sentiment Analysis
Oved et al. (2020)	5,226	Interviews with NBA Players	Player Performance Analysis
Bera et al. (2023)	N/A	Audio Recordings and Past Performance Metrics	Player Performance Analysis
Velichkov et al. (2019)	N/A	Interviews with Athletes	Game Outcome Prediction
Multimodal Dataset			
Sanabria et al. (2021)	N/A	Human Observers in Matches	Summarization
Qiu et al. (2024)	5,100	Videos from YouTube, Annotations by Human Experts	Summarization
Khan et al. (2020)	104	Multiple Video-sharing Platforms	Highlight generation
Narwal et al. (2023)	9514	Cricket Tournaments	Video Segmentation+Audio Classification
Jardim et al. (2023)	More than 1.5 million	Fandom sports wikis	Open-ended-QA
bench authors (2023)	986	Generated by the Author	Classification
Xia et al. (2024)	70,592	Existing QA datasets	Multiple-Choice-QA
Convertible Datasets			
Xia et al. (2022)	1,632	NCAA men's Volleyball Games	Prediction and Analyzation
Li et al. (2021)	3,200	Multiple Competition Records	Classification
Liu et al. (2020b)	1,484	Figure Skating Championships	Multiple tasks
Shao et al. (2020)	N/A	Gymnastics Competitions	Multiple Tasks

Table 1: Overview of Language-Based, Multimodal, and Convertible Datasets for Sports Applications. N/A indicates that the specific data size was not mentioned in the original papers.