### **Lit Review**

The literature on sports video summarization and event detection demonstrates a variety of approaches leveraging audio, visual, and textual modalities.

Paper [1] proposes a multimodal architecture for cricket highlight generation using YOLO for segmenting video based on the bowler’s position and a fine-tuned BERT model to classify events from commentary. Their approach achieves a very high accuracy (97%) in classifying events due to a comprehensive list of event-specific word corpus. The temporal misalignment between commentary and video introduces errors in classification.

Paper [2] introduces a weighted dynamic heartbeat graph in order to identify events from Twitter streams. A temporal graph is constructed from the tweet content, and a rule-based classifier identifies event candidates based on graph features. This method has been validated on benchmarks like the FA Cup, and demonstrates robust event detection.

Paper [3] focuses on automatic cricket video summarization using optical character recognition (OCR) and audio-based event detection. The combination of CNNs and OCR achieves higher accuracy than other methods, effectively identifying events like boundaries and wickets. Limitations include challenges in segmenting fast-paced videos and variations in scoreboard design.

Paper [4] uses audio-based features for near real-time event detection in soccer broadcasts. Block energy and acoustic repetition indices are used to identify key events such as goals. The system achieves effective summary generation but relies only on audio cues, limiting its multimodal capabilities.

While these approaches offer valuable insights, our work addresses gaps in multimodal integration for football video summarization. Unlike [1] and [3], which focus on cricket, we incorporate football-specific nuances, such as integrating commentary, scoreboard detection, and Twitter data to enhance recall and enrich summaries. Furthermore, we overcome limitations in [4] by combining audio and visual modalities to improve detection reliability and context.

### 

### **Commentary Classification Using BERT**

In the proposed multi-modal summarization approach, we fine tune a BERT (Bidirectional Encoder Representations from Transformers) model to classify football commentary text into specific events in football. This modality makes use of BERT’s natural language understanding to recognise and classify specific patterns in football commentary, and map each commentary segment to a specific event category such as "free kick," "foul," or "attempt."

The event classification process begins with us making use of the BERT model to obtain dense vector textual representation for the commentary text. Each commentary segment is cut off to 512 tokens because that is the maximum context window size in BERT and it is also long enough such that we do not lose relevant context around the specific event. This ensures that the model processes the entire necessary input without any loss due to truncation, therefore capturing unique patterns in commentary that will help in effectively classifying football events.

During training, we trained the BERT model for multi-class classification with 5 training epochs, achieving a final classification accuracy of 98%.

The fine-tuning process was done by utilizing the **TRAINER** API from the **TRANSFORMERS** library, which manages training, validation, and checkpointing:

Code:

from transformers import Trainer, TrainingArguments

training\_args = TrainingArguments(

output\_dir="./bert\_results",

num\_train\_epochs=5,

per\_device\_eval\_batch\_size=16,

eval\_strategy="epoch",

logging\_dir="./bert\_logs"

)

trainer = Trainer(

model=model,

args=training\_args,

train\_dataset=train\_dataset,

eval\_dataset=val\_dataset,

compute\_metrics=compute\_metrics

)

This commentary-based classification helps our system to accurately recognize and prioritize significant moments by using commentary (text modality), enhancing the multi-modal summarization framework. This approach shows how LLM based fine-tuning can be used to achieve high precision for real-world sports event classification tasks.

**Multi-modal Integration**

In the proposed framework for multi-modal sports video summarization, integrating data from various different sources is of utmost importance to accurately identify and prioritize significant events. For this we propose a timestamp clustering and weighting algorithm that synthesizes temporal data from all the modalities (commentary, crowd noise, score board).

Let **Ti,j** ​represent the **jth** timestamp of the **ith** modality, where **i ∈ {1,2,…,M}** for each of the **M** modalities, and **j ∈ {1,2,…,Ni}** for **Ni**​ timestamps from the **ith** modality.

To determine overlapping timestamps across multiple modalities, we define a **clustering window of** **Δt seconds** that groups timestamps into clusters if they fall within **Δt** seconds of each other. Let **Ck**​ denote the **kth** cluster, which includes timestamps **{Ti,j}** such that:

**∣Tij−Ti′j′∣ ≤ Δt, ∀ i,i′ ∈{1,2,…,M},j,j′∈{1,2,…,Ni.**

Each cluster **Ck** is calculated a **weight** **wk**​ based on the number of modalities contributing to that cluster:

**wk =∣ {i Ti,j ∈ Ck} ∣**

Thus, the weight **wk**​ signifies the significance of the cluster, with higher values indicating more importance because of multi-modal consensus on the event's importance.

For customisable durations of summaries, clusters **Ck** are sorted in descending order by weight **wk**, giving higher priority to events that have a multi-modal consensus. Based on the desired length of summary (short, medium, long), a subset of the top clusters is selected to fit the target duration. Thus, the number of clusters selected **Nselected** varies dynamically to produce a summary of the desired length.

**Nselected = Top clusters sufficient to match the target summary duration**

This selection mechanism allows the algorithm to effectively integrate the timestamps generated by each modality into a single summary of desired length.

### **Results**

This section shows the performance of our proposed multi-modal system for sports video summarization showcasing how it can incorporate information from Tweets, score board detection and audio analysis.

**Events Detection from Twitter Data Result**

The twitter-based event detection was successfully able to identify significant events by the peaks in the heartbeat score. The keyword corresponding to each peak was identified in order to contextualize each peak. However, noise in the form of occasional fluctuations in tweet activity resulted in a few missed classifications.

#### **Scoreboard Extraction and Detection Result**

YOLOv4 successfully detected scoreboards in different videos, as well as successfully extracted essential information (team names, scores, time) from high-resolution frames (1280×720, 30 frames per second) with an accuracy of 89.06 percent. Sometimes low-resolution frames and noisy frames negatively impacted the accuracy for the algorithm.

**Audio Analysis and Commentary Classification Result**

The audio analysis module, using Whisper for transcription, managed to attain 95% transcription accuracy. Event detection based on RMS intensity and peak finding was generally accurate in detecting actual key events; however, sometimes the timestamp detection was noisy, especially from suboptimal quality audio inputs. While using BERT to classify the commentary, the model was able to get over 95% accurate classifications. It was understood that certain misclassifications occurred due to the differences between commentary and the actual events taking place.  
We chose **Goal & Scoring Ratio** (Goals and scoring opportunities detected by our model / Total goals and scoring opportunities in the match) as the primary metric for evaluating the audio modality, as it directly measures the effectiveness of capturing goal and near goal events. Accuracy was not selected as a metric because most of the moments in a football match do not involve key events like goals. A model could get a high accuracy just for predicting ‘no event’ for most of the match thus not testing its capability of easily finding important highlights. On the other hand, Precision and Recall are more suitable metrics in evaluating the models capabilities of correctly identifying relevant important events.

| **Match** | **Goals and scoring opportunities Detected by Model** | **Total Goals and scoring opportunities** | **Goal Ratio (%)** |
| --- | --- | --- | --- |
| France vs. Croatia | 10 | 15 | 66% |
| Portugal vs. Spain | 8 | 12 | 66.6% |
| Belgium vs. Brazil | 10 | 16 | 62.5% |

The audio modality's **Goal Ratio** indicates its ability to capture goals based on sound intensity. While it missed some goals, it consistently detected key moments, demonstrating its importance in event detection.

#### **Multi-modal Integration and Comparative Analysis**

In order to evaluate our overall model, we compare its output with popular highlights available on YouTube for that match, as these summaries are curated to showcase key moments of a match and serve as a widely accepted benchmark for event detection quality.

Our system-generated summaries captured all events included in popular YouTube highlights and additional near-goal events not present in the highlights. While these extra events might be considered false positives, they represent key moments that enhance the viewer's experience, making our summaries more enriching.

**Metric Definitions**:

* **Precision**: Relevant events captured by the model / Total events captured by the model.
* **Recall**: Relevant events captured by the model / Total relevant events in the YouTube summary.

**Performance Metrics for Multi-modal Integration**:  
A comparison between events captured by our system and YouTube highlights is summarized below:

| **Match** | **Events in YouTube Summary** | **Events in Generated Summary** | **Relevant Events In Detected Highlights** | **Precision** | **Recall (%)** |
| --- | --- | --- | --- | --- | --- |
| France vs. Croatia | 7 | 10 | 9 | 70% | 90% |
| Portugal vs. Spain | 8 | 11 | 9 | 72% | 81.8% |
| Belgium vs. Brazil | 6 | 9 | 7 | 66.6% | 77% |

**Goal Detection in Overall System**:  
The scoreboard detection modality ensured all goals were successfully identified, yielding a **100% Goal Ratio** across matches. This result highlights the crucial role of scoreboard detection in augmenting the system's accuracy for key event identification.

| **Match** | **Goals Detected by Model** | **Total Goals** | **Goal Ratio (%)** |
| --- | --- | --- | --- |
| France vs. Croatia | 6 | 6 | 100% |
| Portugal vs. Spain | 6 | 6 | 100% |
| Belgium vs. Brazil | 3 | 3 | 100% |

By integrating scoreboard detection, our system ensures complete coverage of goals, addressing the limitations of audio-based detection. This proposed multi-modal approach provides a more reliable summarization of critical moments in football matches.