Text Summarization of English Premier League Soccer Matches

p-20

Terry Quach

Banner ID: B00919525

Prof. Dr. Vlado Keselj

CSCI 4152

2026-11-8

Table of Contents

Problem Statement

In the digital era, soccer fans' access to quick and informative updates about matchdays is often limited to basic scorelines or word of mouth. For those who wish to understand what truly happened during a game, long-form match reports remain the primary source of information. However, these reports are often verbose, unstructured, and time-consuming to read, which makes it difficult for fans to stay informed efficiently.

The goal of this project is to design and implement a Natural Language Processing (NLP) system that extracts and summarizes essential information from English Premier League (EPL) match reports. The system will process unstructured text from multiple sources to generate concise summaries that highlight key match events, including the final score, goal scorers, bookings, penalties, and other notable moments.

This project aims to address two main problems: first, the lack of engaging, accessible information for casual fans, and second, the overload of lengthy, detailed reports that discourage quick consumption. By enabling fast, accurate, and structured representations of match data, the proposed system will allow users to obtain meaningful "at-a-glance" updates on games. Furthermore, it will benefit sports journalists by serving as a bridge between their in-depth analyses and a broader audience.

Recent research has shown growing interest in automating the summarization of sports games using NLP techniques. For example, Wang et al., 2021 introduced the *K-SportsSum* dataset, which pairs live commentaries with news articles to generate high-quality sports summaries. Their work highlights two key challenges in sports summarization: (1) the presence of noisy and unstructured data in automatically collected commentary–news pairs, and (2) a knowledge gap between informal, event-driven commentaries and professionally written sports articles. This project extends the concept to the EPL context.

The goal is to create a scalable system that converts unstructured match reports into concise, human-readable summaries.

1 Possible Approaches

This project combines multiple NLP techniques for information extraction, event recognition and text summarization. The system will be developed in modular stages to handle web scraping, pre-processing, extraction and summarization.

1.1 Data Collection and Pre-processing

Match reports, statistics, and live commentary data will be collected from the official Premier League website and Google's live commentary panel using **Scrapy**, a Python-based web scraping framework. The retrieved text will then undergo several pre-processing steps using **spaCy**, including tokenization, stop-word removal, and normalization, to clean and prepare the data for analysis. Finally, the processed text will be converted into structured `JSON` format to facilitate subsequent summarization and named entity recognition tasks.

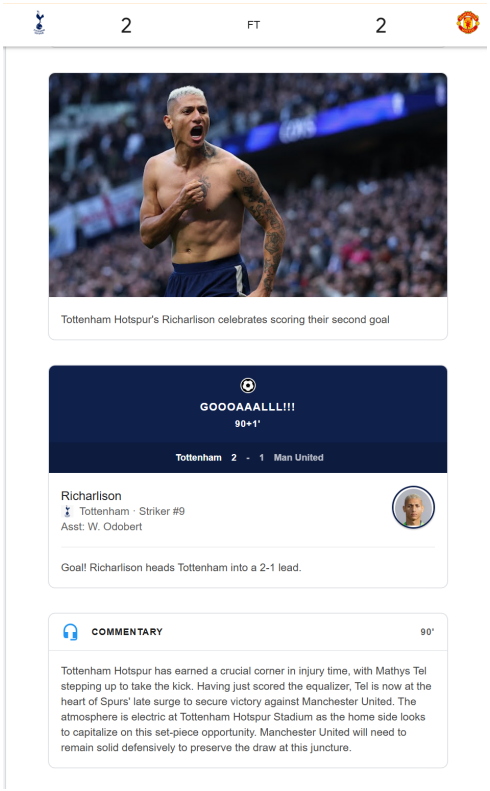

Figure 1: Sample of Google's live commentary from the TOT - MUN soccer match on Nov. 8, 2026 Google LLC, 2025

1.2 Named Entity Recognition (NER)

Named Entity Recognition (NER) will be employed to identify and categorize key entities such as player names, team names, venues, and referees within the collected match reports. The system will utilize transformer-based models available in **spaCy** to detect entities with high precision. To improve domain-specific accuracy, the NER model may be fine-tuned on football-related text data. Post-processing rules will be implemented to standardize variations in entity names (for example, mapping "Man City" to

"Manchester City") to ensure consistency across summaries.

1.3 Rule-Based and Regex Event Extraction

Following entity recognition, a rule-based extraction component will identify structured match events such as goals, bookings, substitutions, and penalties. Regular expressions and dependency parsing techniques provided by **spaCy** will be used to extract temporal and contextual information, such as the minute of play and the player involved in each event. For example, a phrase like "*90+1' – GOAL – Richarlison*" will be parsed to extract the event type (goal), player (Richarlison), and timestamp (90+1th minute). Extracted events will be represented in structured formats such as `JSON` objects to facilitate downstream summarization and visualization.

1.4 Event and Relationship Modeling

Once events are extracted, they will be represented in a structured format to capture the relationships between entities, actions, and outcomes. Each match will be modeled as a collection of event objects containing attributes such as event type, player, team, and time. This structured representation enables downstream summarization and analytical operations.

For instance, extracted information will be organized into a `JSON` schema similar to below:

```
{
  "match": "Arsenal vs Chelsea",
  "score": "1-0",
  "events": [
    {"minute": 23, "type": "goal", "player": "Bukayo Saka",
        "team": "Arsenal"},
    {"minute": 44, "type": "yellow card", "player": "Enzo Fernández",
        "team": "Chelsea"}
  ]
}
```

This structured approach allows for better traceability of entities and supports efficient data querying, aggregation, and visualization.

1.5 Text Summarization Techniques

The summarization component will generate concise, human-readable overviews of each match based on the extracted structured data. Two primary approaches will be explored: extractive and abstractive summarization.

In the **extractive** approach, **TextRank** will be employed to identify and select the most informative sentences from the original reports. This method ensures factual accuracy while condensing lengthy descriptions into a compact summary.

For the **abstractive** approach, transformer-based models such as **BART** and **T5** will be utilized to generate more natural, paraphrased summaries. These models are capable of understanding context and rephrasing text to produce coherent narratives, for example: "*Tottenham vs United was a back-and-forth match, as extra-time goals leave both clubs with 1 point each*"

Both methods will be evaluated for clarity, informativeness, and fluency to determine which approach best suits the intended use case.

1.6 Evaluation and Metrics

The system's performance will be evaluated across two main components: information extraction and text summarization.

For **information extraction**, evaluation will focus on the accuracy of identified entities and events using standard metrics such as precision, recall, and F1-score. A small manually annotated dataset of match reports will serve as the ground truth for comparison.

For **summarization**, quality will be assessed using automatic evaluation metrics such as *ROUGE* or *BLEU*, which measure content overlap between generated summaries and human-written references. Additionally, a qualitative evaluation will be performed to judge readability, coherence, and perceived informativeness.

Qualitatively, this project adopts the evaluation framework proposed by Moon (2009), who assessed the quality of football video summaries using the "four C's" criteria: *Coverage*, *Conciseness*, *Context*, and *Coherence*. These principles can be adapted to textual summarization as follows:

- **Coverage** – The summary should include all key match events such as goals, bookings, and turning points.
- **Conciseness** – The generated summary should be short and to the point, avoiding unnecessary repetition or verbose commentary.
- **Context** – Each event should be presented with enough background for the reader to understand its significance in the match narrative.
- **Coherence** – The summary should flow logically, maintaining narrative consistency between sequential match events.

In addition, user evaluation will be conducted to assess readability, informativeness, and user engagement, particularly among casual fans. This aligns with Moon's interactive summarization goals of providing flexible, individualized sports content that maintains narrative fidelity while adapting to viewer needs (Moon, 2009).

2 Project Plan

TThe focus will be on building a working prototype that can collect, extract, and summarize English Premier League match data in a clear and reliable way. The plan below outlines the main stages of work and their expected order.

2.1 Overview

The work will proceed in four main stages:

1. **Data Collection and Pre-processing** – Gather match reports and commentary from the Premier League website and clean the text using Python tools such as Scrapy and spaCy.
2. **Entity and Event Extraction** – Identify players, teams, and match events (goals, cards, substitutions) using NER models and rule-based patterns.
3. **Summarization Module** – Implement extractive (TextRank) and abstractive (BART or T5) summarization methods to generate concise match summaries.
4. **Evaluation and Refinement** – Test the accuracy of extraction and readability of summaries using small-scale manual checks and simple metrics like ROUGE.

2.2 Tentative Timeline

Phase	**Tasks and Expected Duration**
Week 1	Set up data collection and preprocessing pipeline. Verify scraping and text cleaning.
Week 2	Implement NER and event extraction, test on sample matches.
Week 3-4	Develop and compare extractive vs. abstractive summarization.
Week 5	Evaluate summaries, refine results, and finalize documentation.

2.3 Expected Outcomes

By the end of the project, the system should:

- Collect and preprocess EPL match text data.
- Extract structured match events such as goals and bookings.
- Generate short, readable match summaries.
- Provide basic evaluation results showing summary quality and accuracy.

2.3.1 Sample Match Summarization Output

The following example illustrates the expected output of the proposed summarization system. It demonstrates how extracted match events and narrative flow can be synthesized into a concise, coherent summary:

> *Arsenal edged past Chelsea 2–1 in a thrilling London derby at the Emirates. Bukayo Saka opened the scoring in the 23rd minute with a composed finish after a quick counterattack, but Enzo Fernández leveled for Chelsea just before halftime. The second half saw Arsenal dominate possession, and their persistence paid off when Martin Ødegaard curled in the winner from the edge of the box in the 78th minute. Despite late pressure from Chelsea, the Gunners held firm to secure three crucial points and extend their unbeaten run at home.*

References

Google LLC. (2025, November 8). *Search results for "tottenham vs manchester united live commentary"* [Accessed: 2025-11-08]. https://www.google.com/search?num=12&sca_esv=ea796d5874e27d32&rlz=1C1GEWG_enCA1182CA1183&sxsrf=AE3TifPx0QrYXgERLGmLOu_8GayPFxeEHg:1762640303516&q=tot+mun&source=lnms

Moon, B. B. (2009). *Interactive football summarization* [Master's thesis, Brigham Young University] [Master's Thesis]. Retrieved November 8, 2025, from https://scholarsarchive.byu.edu/etd/1999

Wang, J., Li, Z., Zhang, T., Zheng, D., Qu, J., Liu, A., Zhao, L., & Chen, Z. (2021). Knowledge-enhanced sports game summarization. *Proceedings of the 15th ACM International Conference on Web Search and Data Mining (WSDM '22)*, 1165–1173. https://doi.org/10.1145/3488560.3498405