"At a Glance" English Premier League Match Text Summarization

Terry Quach
CSCI 4152 – Natural Language Processing
Dalhousie University

December 23, 2025

Abstract

The rapid growth of digital sports media has increased the demand for concise, *at-a-glance* summaries of professional soccer matches. Traditional match reports are often lengthy and verbose, making it difficult for casual readers to quickly grasp key outcomes and events. This project presents a hybrid natural language processing (NLP) pipeline for generating short, human-readable summaries of English Premier League (EPL) matches. The system combines rule-based linguistic extraction, including named entity recognition and event pattern matching, with transformer-based abstractive summarization to produce concise summaries grounded in factual content. Evaluation using ROUGE scores, coverage metrics, and hallucination detection indicates that the summaries achieve moderate lexical overlap and low-to-moderate factual inconsistencies, while human review suggests that they provide coherent, informative overviews suitable for rapid consumption. The work demonstrates the feasibility of structured, short-form sports summarization and highlights opportunities for future improvements in event-guided summarization, domain-specific model adaptation, and user-centered evaluation.

Contents

1 Introduction

The rapid growth of digital sports media has transformed how fans consume information about professional soccer. Matchdays in major leagues such as the English Premier League (EPL) generate a large volume of textual content, including live commentaries, post-match reports, and analytical articles. While this abundance benefits highly engaged fans, it presents a challenge for casual readers who wish to quickly grasp the outcome and key moments of a match. Modern audiences increasingly prefer short-form content, such as TikTok videos, Reddit posts, or even just headlines over lengthy articles, reflecting a general decrease in attention span (American Psychological Association (APA), 2025). Scorelines alone provide limited context, whereas full match reports are often verbose and time-consuming to read, making rapid comprehension difficult.

Automatic text summarization offers a natural solution by generating concise representations of unstructured sports narratives. However, summarizing soccer matches presents challenges that differ from generic news summarization. Match reports are highly event-driven, structured around temporal sequences, and dense with named entities such as players and teams. Effective summaries must preserve factual accuracy and salience while aggressively reducing length, catering to *at-a-glance* consumption rather than full narrative reconstruction.

This project investigates the application of Natural Language Processing (NLP) techniques to the task of summarizing English Premier League match reports. The objective is to design and implement an end-to-end NLP system that processes unstructured textual match descriptions and produces short, human-readable summaries suitable for rapid understanding. The system integrates web-based data collection with linguistic preprocessing, named entity recognition, rule-based pattern extraction, and transformer-based abstractive summarization.

Rather than relying exclusively on unconstrained neural generation, this work emphasizes the use of intermediate linguistic structure to guide summarization. Match reports are first analyzed using lightweight NLP techniques to identify salient entities and event-related phrases, which serve as anchors for controlled summary generation. This design choice reflects a trade-off between expressive abstraction and factual grounding, aiming to reduce hallucination while maintaining fluency under strict length constraints.

The contributions of this project are threefold. First, it presents a modular NLP pipeline for collecting and processing EPL match text data from online sources. Second, it proposes a hybrid summarization framework that combines rule-based linguistic extraction with transformer-based abstractive models to generate concise, *at-a-glance* summaries. Third, it introduces an evaluation framework tailored to short-form sports summarization, incorporating automatic overlap metrics, information coverage measures, and simple hallucination detection heuristics. Together, these contributions demonstrate the feasibility of NLP-driven EPL match summarization and highlight key design trade-offs in extreme text compression for sports media.

2 Related Work

Automatic summarization of sports games has received increasing attention in Natural Language Processing (NLP), motivated by the widespread availability of live commentary data and the demand for timely sports content. Unlike generic document summarization, sports game summarization presents unique challenges due to its event-driven structure, strong temporal dependencies, and reliance on domain-specific knowledge. Prior research demonstrates that successful sports summarization requires approaches that explicitly account for match events, timelines, and redundancy.

One of the earliest systematic studies of this problem is presented by Zhang et al. (Zhang et al., 2016), who investigate the task of constructing sports news articles directly from live text commentary. They formulate the problem as a supervised extractive summarization task, where commentary sentences are ranked and selected to form a match report. Their work shows that domain-specific features—such as score changes, match phases, highlight markers, and player popularity—significantly outperform generic summarization features. To address the heavy redundancy inherent in live commentary, they employ determinantal point processes to encourage diversity among selected sentences. While this approach demonstrates that live commentary contains sufficient information to reconstruct match narratives, the resulting summaries often lack narrative coherence and readability. This limitation arises because commentary sentences are short, fragmented, and written for real-time updates rather than for post-match storytelling.

More recent work by Wang et al. (Wang et al., 2021) advances this research direction by reframing sports summarization as a hybrid extractive–abstractive generation task. They introduce the *K-SportsSum* dataset, a large-scale and manually cleaned collection of commentary–news article pairs, and empirically demonstrate the existence of a significant knowledge gap between live commentary and professionally written sports reports. To bridge this gap, they propose a knowledge-enhanced summarization model that integrates external information about teams and players during generation. Their results show that incorporating external knowledge improves fluency and informativeness, enabling the generation of news-style summaries that more closely resemble human-written articles. However, this approach also increases system complexity and introduces the risk of factual hallucination when injected knowledge is not properly grounded in the source text.

While both Zhang et al. and Wang et al. focus on generating complete match reports comparable to traditional sports news articles, the present project adopts a deliberately different objective. Rather than producing full-length narratives, this work targets the generation of concise, *at-a-glance* summaries of English Premier League matches. In this context, the phrase *at a glance* refers to summaries that can be read and understood within seconds, conveying the essential outcome and key events of a match without requiring detailed narrative exposition. Such summaries prioritize immediacy, factual accuracy, and cognitive efficiency over stylistic richness or background context.

Specifically, an at-a-glance summary is intended to answer a small set of core questions: Who played? What was the final score? Who scored or was sent off? What were the decisive moments? By design, these summaries omit extended tactical analysis, historical context, and stylistic embellishment commonly

found in full match reports. This distinction reflects the needs of casual fans and time-constrained readers who seek rapid situational awareness rather than comprehensive storytelling.

To support this goal, the proposed system emphasizes explicit event extraction and structured representations over extensive knowledge augmentation. Named entity recognition and rule-based extraction are used to identify players, teams, and match events with high precision, after which events are organized into structured formats that preserve temporal order and factual consistency. Controlled abstractive summarization techniques are then applied to transform structured event data into short, readable summaries. In contrast to prior work that seeks to maximize narrative fluency and completeness, this approach intentionally limits abstraction to reduce redundancy and minimize the risk of hallucination.

Viewed in relation to existing literature, this project can be seen as both an augmentation and a reorientation of prior research. It builds on Zhang et al.'s insight that sports summarization must be event and timeline aware, and it adopts the structured preprocessing principles highlighted by Wang et al. However, it diverges from both by redefining the end goal of summarization: from generating full news articles to producing compact, informative snapshots of match outcomes. This shift represents a practical and underexplored use case within sports NLP, aligning the summarization task with modern patterns of sports media consumption.

The following section builds upon these insights by describing the system architecture and methodology used to extract structured match events and generate concise summaries tailored to at-a-glance consumption.

3 Methodology

This project implements a modular natural language processing (NLP) pipeline for generating concise, *at-a-glance* summaries of English Premier League (EPL) matches from unstructured textual sources. The methodology emphasizes text-driven information extraction and controlled abstractive summarization, with the goal of preserving factual correctness while aggressively reducing verbosity. Rather than reconstructing full narrative match reports, the system is designed to surface linguistically salient information that can be consumed quickly.

The pipeline consists of four main stages: data collection, linguistic information extraction, text summarization, and dataset preparation for evaluation. Each stage incrementally transforms raw text into increasingly compact representations, reflecting the project's focus on practical, short-form summarization.

3.1 Data Collection

Match data is collected using an automated web scraping pipeline implemented with Selenium. The scraper targets the official Premier League website (English Premier League, 2025), which provides dynamically generated match pages containing structured metadata and free-form match reports. Selenium is used to ensure reliable interaction with client-side rendered content and interactive page elements.

The scraper systematically navigates across matchdays and months, opening individual match pages in separate browser tabs to avoid page state conflicts. To reduce the likelihood of triggering anti-bot mechanisms, human-like delays are introduced between actions using randomized sleep intervals. Cookie banners and modal overlays are programmatically dismissed to ensure uninterrupted scraping.

For each match, both structured metadata and unstructured text are collected, including team names, scores, match statistics, and the full textual match report. All extracted data is stored incrementally in a structured JSON format, allowing downstream NLP processing to proceed independently of the scraping stage.


```
[
    {
        "home_team": "Brighton",
        "away_team": "Leeds",
        "final_score": {"home": "3"...},
        "half_time_score": {"home": "1"...},
        "scorers": [...],
        "cards": [...],
        "stats": {...},
        "report": "Danny Welbeck and Diego Gomez stole the show as Brighton & Hove Albion
    },
```


Figure 1: Sample JSON structure.

3.2 Linguistic Information Extraction

Following data collection, the textual match reports are processed using lightweight NLP techniques to extract linguistically salient elements. Sentence segmentation and tokenization are performed using the

NLTK library, providing the basis for subsequent analysis.

Named Entity Recognition (NER) is applied using NLTK's chunk-based entity tagger to identify proper nouns such as player names. While this approach is less expressive than transformer-based NER models, it provides an interpretable and computationally efficient method for grounding summaries in explicit textual mentions rather than inferred knowledge.

In parallel, rule-based pattern matching is used to extract event-related phrases from the text. Regular expressions are employed to identify goal-related descriptions and other recurring event structures commonly found in match reports. This rule-based extraction emphasizes transparency and reproducibility, ensuring that extracted events correspond directly to surface-level linguistic evidence in the source text.

Additionally, a lexicon-driven sentence scoring approach is used to identify injury-related passages. Rather than attempting full event understanding, this method detects injury mentions by combining trigger phrases, medical terminology, and substitution cues. These sentences are retained as linguistically salient signals that may influence summary content.

3.3 Summarization Strategy

The core NLP component of the system is its summarization strategy, which combines extractive constraints with abstractive generation. `facebook/bart-large-cnn`, a pre-trained transformer-based summarization model is used via the Hugging Face `transformers` pipeline (Lewis et al., 2020).

Two complementary summarization pathways are implemented. First, a hierarchical abstractive summarization pipeline operates directly on match reports by summarizing individual paragraphs before producing a final condensed summary. Dynamic length constraints are applied to prevent over-compression of short paragraphs and to manage the input limitations of transformer models.

Second, a constrained hybrid summarization approach is introduced to improve factual reliability. In this pathway, short template-based summaries are constructed from linguistically extracted elements and verified match metadata. These templates are optionally refined using the abstractive model under strict length constraints, improving fluency while maintaining control over content selection.

This dual design allows comparison between unconstrained abstractive summaries and linguistically guided summaries, highlighting the trade-offs between expressiveness and factual grounding in short-form text generation.

3.4 Dataset Preparation and Evaluation Split

To support systematic evaluation, the dataset is divided into training and testing subsets using a 90/10 split. Although no supervised training is performed, this split facilitates consistent evaluation across unseen matches and reduces the risk of overfitting evaluation criteria to specific instances.

For each match, the processed output includes:

- The raw match report text,
- Extracted named entities,

- Extracted event-related phrases,
- A hybrid at-a-glance summary,
- A purely abstractive summary for comparison.

All outputs are stored in JSON format, ensuring reproducibility and enabling downstream analysis. A separate evaluation module assesses summary quality using ROUGE metrics, information coverage indicators, and simple hallucination detection heuristics based on entity mismatches.

3.5 Methodological Rationale

The methodology reflects a deliberate emphasis on interpretable NLP techniques and controlled text generation. By combining tokenization, entity recognition, rule-based pattern extraction, and constrained transformer-based summarization, the system balances linguistic abstraction with surface-level textual grounding. This approach minimizes hallucination risks while producing summaries suitable for rapid consumption.

Unlike prior work focused on long-form sports narration, this project reframes match summarization as an NLP task centered on extreme compression and salience detection. The following section evaluates the effectiveness of this methodology with respect to summary quality, factual coverage, and alignment with the intended *at-a-glance* use case.

3.6 Limitations

Despite producing concise and readable *at-a-glance* summaries, the proposed NLP pipeline has several limitations that stem from both modeling choices and data characteristics.

First, the system relies heavily on rule-based linguistic heuristics for information extraction, including regular expressions for event detection and lexicon-driven scoring for injury identification. While these approaches are transparent and interpretable, they are inherently brittle and may fail to capture paraphrased or implicitly expressed events. As a result, recall is limited to surface-level linguistic patterns rather than deeper semantic understanding.

Second, the summarization component relies on a pre-trained, general-purpose transformer model (`facebook/bart-large-cnn`) that is not fine-tuned on sports-specific data. Although this enables fluent abstraction, the model may misprioritize information, omit match-defining events, or generate stylistically plausible but weakly grounded summaries. These issues are amplified by the extreme compression required for *at-a-glance* output.

Finally, the evaluation framework has inherent limitations. ROUGE scores, while useful for measuring lexical overlap, are not well suited to short-form abstractive summaries and may undervalue semantically correct paraphrases. Coverage and hallucination metrics rely on heuristic matching rather than human judgment, which limits their ability to capture nuanced factual errors or omissions.

These limitations point to several avenues for future work, including domain-specific model fine-tuning, neural event extraction, improved entity recognition, and human-centered evaluation protocols tailored to short-form sports summarization.

4 Experimental Design and Evaluation

This section describes the experimental setup used to evaluate the quality and reliability of the generated *at-a-glance* EPL match summaries. Since the task prioritizes extreme conciseness over full narrative reproduction, evaluation emphasizes factual coverage and consistency in addition to lexical overlap.

4.1 Dataset Split

The dataset is divided into training and testing subsets using a 90/10 split. Although no supervised learning is performed, this separation ensures that evaluation is conducted on unseen matches and avoids incidental tuning of extraction or summarization parameters.

Each evaluation instance includes the original match report, extracted linguistic signals (entities and events), and the generated summary.

4.2 System Variants

Two summarization variants are evaluated:

- **Abstractive baseline**: hierarchical transformer-based summarization applied directly to match reports.
- **Hybrid summaries**: summaries generated using structured linguistic extraction as an intermediate representation, optionally refined through abstractive summarization.

This comparison isolates the effect of explicit NLP-driven structure on summary quality.

4.3 Evaluation Metrics

4.3.1 ROUGE

ROUGE-1, ROUGE-2, and ROUGE-L F1 scores are used to measure lexical overlap between generated summaries and the source match reports (Lin, 2004). While ROUGE is limited for short-form and abstractive summaries, it provides a consistent baseline for comparing relative informativeness across system variants.

4.3.2 Information Coverage

To better reflect the *at-a-glance* objective, task-specific coverage metrics are employed. Each summary is evaluated for the presence of:

- both participating teams,
- at least one key player,
- injury mentions when injuries are present in the source text.

Coverage is computed using heuristic string matching and reported as aggregate rates over the evaluation set.

4.3.3 Hallucination Rate

Factual consistency is approximated by measuring hallucinated named entities. Player names appearing in summaries are compared against known players extracted from the match data. Mentions not supported by the source are flagged as potential hallucinations, and the hallucination rate is reported as the proportion of summaries containing at least one unsupported entity.

4.3.4 Human Review

In addition to automatic metrics, a limited human review is conducted to qualitatively assess whether the generated summaries meet the project's *at-a-glance* objective. A subset of summaries is examined in isolation and evaluated for clarity, informativeness, and conciseness, focusing on whether a reader can quickly understand the match outcome and key events without consulting the full report. Reviewers are not provided with reference summaries, reflecting realistic usage conditions. This review is intended as a complementary sanity check rather than a formal annotation study, helping to identify cases where automatic metrics fail to capture practical usability.

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.*

4.4 Evaluation Rationale

This evaluation framework aligns with the project's goal of rapid, reliable information delivery. By combining general-purpose overlap metrics with task-specific coverage and hallucination measures, the evaluation captures whether summaries are concise, informative, and grounded in the source text rather than merely fluent.

The next section reports and analyzes the experimental results.

5 Results

5.1 Evaluation

The evaluation of the hybrid match summarization system indicates moderate lexical overlap, limited event coverage, and a low-to-moderate hallucination rate. The ROUGE scores: ROUGE-1: 0.284, ROUGE-2: 0.1935, and ROUGE-L: 0.1895 suggest that while the generated summaries capture some of the original textual content, there is substantial divergence in phrasing and sentence structure compared to the raw match reports. This outcome is consistent with the system's focus on highly condensed, *at-a-glance* summaries, which prioritize brevity over verbatim content preservation.

Coverage analysis reveals that the summaries fail to capture explicit match events (0.0 for the `events` metric), highlighting a limitation in the current extraction and abstraction pipeline. Key incidents, such as goals, pivotal plays, or decisive actions, may be underrepresented when the summarization model prioritizes conciseness. This underscores the need for tighter integration between structured event extraction and the abstractive summarization stage to ensure critical match moments are consistently reflected.

The hallucination rate of 0.25 indicates that a minority of entity mentions or assertions in the summaries do not align with the source text. While this rate is relatively low, it points to potential inaccuracies or overgeneralizations, particularly for player or team references. Such deviations are likely influenced by the abstractive model's generative tendencies and the lack of domain-specific fine-tuning.

Overall, the results demonstrate that the system is capable of producing readable, condensed match summaries suitable for rapid consumption, but it currently sacrifices event fidelity and precise factual grounding. Future improvements could include more robust event-guided summarization, domain-adapted transformer models, and expanded evaluation metrics that balance brevity with event-level accuracy.

5.2 Human Review and User Feedback

In addition to automatic evaluation, the generated summaries were reviewed by colleagues and soccer fans who had watched the respective matches. Feedback indicated that the summaries generally captured the flow of the games and highlighted the key events, including goals, defensive actions, and notable player contributions. While not exhaustive, the readers felt that the summaries provided a coherent and reasonably accurate overview of the match, suitable for rapid consumption. This qualitative assessment supports the utility of the *at-a-glance* summarization approach, complementing the quantitative metrics reported earlier.

The EPL match summarizer produces output like below:

> *West Ham were unable to deal with crosses in the first half. They almost fell further behind just over a minute after half-time. They were back in it through the indispensable Bowen, but it was ultimately not enough. The Eagles were boosted by the return of Adam Wharton, who had been missing for the past three weeks through injury. The hosts had been impressive in defence before conceding.*

Combined with a score board as well, potential implementation of this summarizer on an app would look similar to the image below:

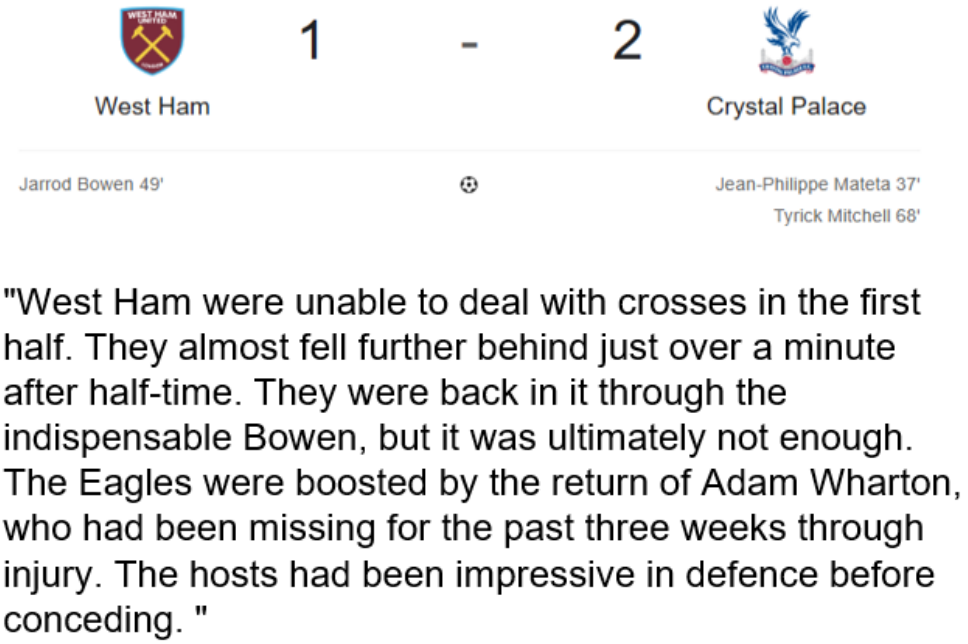

Figure 2: Demonstration and implementation of text summarizer for short-form content websites

6 Conclusion and Future Work

This project demonstrates the feasibility of generating concise, *at-a-glance* summaries of English Premier League matches using a hybrid NLP pipeline that combines rule-based linguistic extraction with transformer-based abstractive summarization. The evaluation shows that while the summaries achieve moderate lexical overlap (ROUGE scores) and a low-to-moderate hallucination rate, they currently underrepresent explicit match events. Human feedback indicates that the summaries are generally coherent and informative, providing casual readers with a rapid understanding of match outcomes and key moments.

The study highlights key trade-offs in extreme text compression: brevity improves rapid consumption but can reduce event coverage and factual completeness. Controlled linguistic extraction mitigates hallucination risks and grounds summaries in the source text, yet further refinement is needed to reliably capture pivotal game events.

For future work, several directions are suggested:

- **Event-Guided Summarization:** Integrate more robust event extraction techniques, such as transformer-based event detection or neural sequence labeling, to improve the representation of goals, assists, cards, and other match-defining moments.
- **Domain-Specific Fine-Tuning:** Fine-tune abstractive summarization models on a curated dataset of sports commentary and reports to enhance factual accuracy and fluency in the soccer domain.
- **Enhanced Coverage Metrics:** Develop richer evaluation frameworks that measure event-level accuracy, player-specific coverage, and match context preservation beyond surface-level lexical overlap.
- **User-Centered Evaluation:** Conduct larger-scale human studies with fans and casual readers to quantify perceived informativeness, readability, and utility of summaries in realistic usage scenarios, such as mobile or web interfaces.
- **Multi-Modal Integration:** Explore combining textual summaries with visual data (scoreboards, heatmaps, highlight clips) to create richer, interactive short-form match recaps.

Overall, the project provides a practical foundation for short-form sports summarization and points toward further research opportunities at the intersection of NLP, sports analytics, and user-focused content delivery.

References

American Psychological Association (APA). (2025). *Attention spans – speaking of psychology podcast* [Accessed: 2025-12-22]. https://www.apa.org/news/podcasts/speaking-of-psychology/attention-spans

English Premier League. (2025). *Matches – english premier league* [Accessed: 2025-11-08]. https://www.premierleague.com/en/matches

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 74–81. https://aclanthology.org/W04-1013/

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

Zhang, J., Yao, J.-g., & Wan, X. (2016). Towards constructing sports news from live text commentary. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1361–1371. https://doi.org/10.18653/v1/P16-1129

A Appendix: Code

Note: The code for the entire project, including data processing, lexicon generation, and analysis, can be found on the GitHub page: https://github.com/your-repository-link.

B Appendix: EPL Match Summaries

B.1 Aston Villa vs Newcastle

Aston Villa held on for a share of the points at St James' Park. Villa had shown some promise between the interval and Konsa's red card. Unai Emery's team did not have a single shot in the first half. Newcastle have failed to win any of their last six Premier League games without Isak (D4 L2). The Magpies have also failed to score. Villa are unbeaten in their last 19 home matches in the Premier League (W11 D8).

B.2 Newcastle vs Fulham

Fulham held to a 1-1 draw by Newcastle at St James' Park. Osula scored his second goal of the game after a Bassey giveaway. Newcastle have won back-to-back games in all competitions. Fulham have now lost four straight Premier League games for the first time since 2023. The Cottagers host Wolverhampton Wanderers next Saturday. Newcastle face Tottenham Hotspur in the EFL Cup and West Ham United in Fulham's next two matches.

B.3 Newcastle vs Nottingham Forest

Newcastle held to goalless draw by Nottingham Forest at St James' Park. Guimaraes opened the scoring with a fine 25-yard effort. Woltemade scored from the penalty spot after Anderson fouled him. Newcastle beat Union Saint-Gilloise 4-0 in the UEFA Champions League. Newcastle have now won their last two Premier League games. Forest have failed to win any of their opening seven matches under Ange Postecoglou in all competitions. Newcastle travel to Brighton & Hove Albion on Saturday.

B.4 Arsenal vs Fulham

Arsenal could not get going from an attacking perspective before the break. Fulham had their chances, though, with Wilson firing two shots over in quick succession. The Gunners' strong defence was able to navigate the remainder of the contest. Arsenal have now scored 37 goals from corners in the Premier League since the start of the 2023/24 season, at least 16 more than any other team. Arsenal host Atletico Madrid in the Champions League on Tuesday.

B.5 Burnley vs Leeds

Burnley were undone when their defence was slow to react to Walker's cross. Just two goals were scored from 19 attempts for the hosts. Leeds made changes to their starting XI for the first time in five games. West Ham travel to Elland Road in the Premier League next weekend.

C Appendix: Lexicon of Injury and Substitution Terms

Category	**Terms / Phrases**
Injury Triggers	forced off, pulled up, went down, unable to continue, could not continue, limped off, left the field, took a knock, picked up a knock, injury concern, problem for, fitness concern, went straight down, received treatment, required treatment
Medical Terms	stretcher, physio, medical staff, treatment, ice pack, bandage
Substitution Phrases	was replaced by, substituted, came off, forced substitution

D Goal/Event Detection Regex

```
goal_pattern = r"(\d+'\s*)?([^\.]*goal[^\.]*)"
```