Cricket Momentum Score from Ball by Ball Commentary

Presented

In Partial Fulfillment of the Requirements
for the Degree of Bachelor of Statistics

Supervised By: Prof. Diganta Mukherjee

Written By: Aryanil Kundu

Date Last Edited: July 30, 2025

Abstract

In cricket analytics, quantifying momentum shifts is critical for understanding match dynamics and supporting real-time decision-making. Traditional metrics, such as the Pressure Index, rely primarily on numerical match statistics (runs, wickets, and required run rate) and may fail to capture the qualitative context that influences a team's perceived advantage. This study proposes a commentary-driven momentum index derived entirely from ball-by-ball commentary data. Using a manually curated set of positive and negative cricket-specific keywords in conjunction with VADER sentiment analysis, each commentary line is assigned a batting / bowling team–aligned sentiment score. To reduce noise and highlight underlying trends, a simple moving average (SMA) is applied over an 18-ball window, yielding a smoothed momentum trajectory. The proposed index is benchmarked against the Pressure Index to assess its ability to reflect match flow. Experimental results on historical match data demonstrate that the commentary-based index captures both major match events (e.g., wickets, boundaries) and subtle pressure situations (e.g., dot-ball sequences, close chances) that may not immediately impact the scorecard. This lightweight, text-driven approach highlights the untapped potential of unstructured commentary data to complement traditional cricket analytics and opens avenues for richer real-time match state estimation.

Acknowledgements

I would like to express my heartfelt gratitude to Prof Diganta Mukherjee for his invaluable guidance and support throughout this project. His expertise and insights into the field have greatly enriched my understanding and have been instrumental in shaping the direction of my research. I am particularly thankful for his encouragement and constructive feedback, which motivated me to strive for excellence. I would also like to extend my appreciation to my friends, who have been a constant source of support and encouragement. Their collaboration and willingness to share ideas and resources made this project a more enjoyable and enriching experience. I am grateful for their companionship and the many discussions that helped refine my thoughts and approaches. Thank you all for your contributions, which have significantly enhanced the quality and outcome of this project.

Contents

List of Tables

List of Figures

Chapter 1

Report

1.1 Introduction

The concept of momentum plays a pivotal role in the dynamics of cricket matches, influencing not only the outcome of games but also team strategies and decision-making. Traditionally, cricket analytics has relied heavily on numerical match statistics such as runs scored, wickets lost, required run rate, and partnerships to quantify the flow of a game. Metrics like the Pressure Index (PI) attempt to capture these dynamics using scorecard-based information, yet they are inherently limited as they do not incorporate the qualitative context of play. Many match-changing moments, such as near-misses, mounting pressure from successive dot balls, or the psychological impact of aggressive batting, may not be adequately reflected in quantitative statistics alone.

Ball-by-ball commentary data offers an underutilized source of information that captures the narrative and emotional tone of a match. Commentary describes not only scoring events but also the atmosphere, player confidence, and pressure situations as they unfold. For instance, a line such as "Ooh, nearly another first-baller as this one jags in wickedly and just misses the off stump" reflects intense pressure on the batter, yet has no direct effect on the scorecard. Sentiment analysis of such text can provide a batting-team–aligned view of match momentum that is otherwise invisible in traditional statistics. Unlike structured data, commentary can reveal subtle shifts in control, for example, an over filled with close chances and tight bowling even if few runs or wickets are recorded.

This work introduces a commentary-driven momentum index that uses sentiment analysis to infer batting momentum in cricket matches. A curated set of cricket-specific positive and negative keywords is combined with VADER sentiment scores to assign each ball commentary a sentiment value from the batting team's perspective. To highlight underlying match trends, the ball-by-ball sentiment is smoothed using a simple moving average. The proposed index is then compared with the Pressure Index (PI) as a benchmark to evaluate its alignment with established match-flow metrics. By relying solely on unstructured commentary text, this approach highlights how qualitative data can complement traditional cricket analytics and provide richer insight into the narrative of a match.

1.2 Data Description

1.2.1 Dataset Source

We use the *Cricket Scorecard and Commentary Dataset* from Kaggle. The dataset aggregates ball-by-ball records and corresponding textual commentary for multiple international and franchise matches sourced from ESPNcricinfo. It contains both structured scorecard information (e.g., runs, balls, innings) and unstructured, natural-language commentary describing each delivery.

1.2.2 Scope and Modality

The dataset comprises two complementary modalities: (i) **structured** match state variables recorded per delivery, and (ii) **unstructured** text commentary per delivery. This dual view allows us to

compute traditional, scorecard-based indicators (e.g., required run rate, pressure proxies) while deriving narrative signals from the commentary via sentiment analysis.

1.2.3 Relevant Fields Used in This Study

For our commentary-driven momentum index and the comparison with a pressure proxy, we rely on the following fields.

Table 1.1: Primary fields utilized in this study.

Field	Type	Description
`Innings`	Categorical	Innings identifier (e.g., "1st innings", "2nd innings").
`Over_number`	Integer	Over index (starting at 1).
`Over_ball`	Integer	Ball index within the over (typically 1–6).
`Over_Ball_new`	String	Derived key: `(Over_number-1).Over_ball` (e.g., 1.1, 1.2, ...).
`Innings_runs`	Integer	Cumulative runs at the end of the delivery.
`Innings_balls`	Integer	Cumulative balls faced in the innings (1,2,3,...).
`Commentary`	String	Long-form textual commentary for the delivery.(e.g.: good delivery to start with, but goes for four...full at off stump, shapes away to find Gayle's outside edge and it runs wide of slip)
`Commentary_short`	String	Short-form textual summary (when available).(e.g.: Willey to Gayle, FOUR runs)

1.2.4 Preprocessing

We perform light preprocessing tailored to sentiment extraction from commentary:

1. Lowercasing all commentary text; optional removal of English stopwords (NLTK).
2. Handling missing or empty commentary lines by skipping or imputing neutral sentiment.
3. Ensuring per-delivery alignment by indexing on `Innings`, `Over_number`, and `Over_ball`.
4. Constructing `Over_Ball_new` (e.g., 1.1, 1.2, ...) for compact visualization along the innings timeline.
5. For PI comparison, fixing `TotalBalls` to the match format (e.g., 120 for T20) and clipping divisions near innings end.

1.3 Methodology

Our methodology combines cricket-specific keyword-based sentiment scoring with lexicon-based sentiment analysis (VADER) to build a commentary-driven batting momentum index. Figure **??** illustrates the overall pipeline, which consists of three major stages: (i) sentiment extraction from commentary, (ii) smoothing to obtain a momentum trajectory, and (iii) comparison with a traditional Pressure Index (PI) benchmark.

1.3.1 Sentiment Extraction from Commentary

We process the `Commentary` field described in Section 1.2 to compute an *adjusted sentiment score* per ball from the batting team's perspective. We first lowercase and lightly normalize the text (e.g., stripping extra punctuation). We then compute the base sentiment using the VADER sentiment analyzer from the NLTK library, which produces a compound polarity score $s_{\text{vader}} \in [-1, 1]$.

Table 1.2: Cricket-specific keywords grouped by polarity and semantic category.

Polarity	Group	Words / Phrases
Positive	Runs (big)	six, maximum, four, boundary
	Runs (small)	runs, single, double, triple
	Shot types	driven, swept, cut, flicked, pulled, lofted, hit hard
	Shot quality	beautifully timed, cracking shot, glorious, perfect placement, pierces the gap, sweet timing
	Milestones	fifty, half-century, hundred, century, maiden hundred, milestone
	Momentum words	accelerating, good partnership, building innings, taking charge, in control, positive intent
	Bowling errors / extras (benefit batting)	no ball, wide, free hit, overthrows, misfield, dropped catch, leg byes, byes, overstep, bonus run
	Commentator praise	excellent shot, top-class, magnificent, sensational, standout
Negative	Wickets	out, bowled, caught, lbw, stumped, hit wicket, run out, edge taken, top edge, chopped on, nicked behind
	Shot mistakes	poor shot, rash shot, unnecessary shot, mistimed, miscued, airborne and gone, bad decision
	Pressure situations	pressure builds, dot balls, struggling, slow scoring, tight over, dry spell, collapse
	Bowling praise (opponent advantage)	brilliant yorker, sharp turn, beaten, unplayable, deadly spell, nagging line, great over
	Injuries / issues	injury, limping, cramp, needs treatment, retired hurt

To align the sentiment with the batting team's advantage/disadvantage, we define curated lists of **positive** and **negative** cricket-specific keywords. For each commentary line, we adjust the polarity as follows:

$$s_{\text{adj}} = \begin{cases} +|s_{\text{vader}}| & \text{if any positive keyword is present,} \\ -|s_{\text{vader}}| & \text{if any negative keyword is present,} \\ 0.1 \times s_{\text{vader}} & \text{if no keyword is matched.} \end{cases} \tag{1.1}$$

This ensures that events clearly beneficial to the batting side (e.g., "four", "six", "free hit") are assigned positive scores, while detrimental events (e.g., "out", "lbw", "dot balls", "collapse") are negative. If no keyword is detected, a small fraction of the raw VADER score is retained to reflect generic positivity/negativity in the commentary tone.

1.3.2 Rolling Momentum Computation

The ball-by-ball adjusted sentiment scores $\{s_{\text{adj}}(i)\}$ can be noisy due to variability in commentary style and single-ball fluctuations. We therefore compute a smoothed trajectory using a simple moving average (SMA) over a fixed window of $w = 18$ balls (three overs):

$$\text{Momentum}(t) = \frac{1}{w} \sum_{i=t-w+1}^{t} s_{\text{adj}}(i), \qquad t \geq w. \tag{1.2}$$

For $t < w$, we use a partial average over available balls. The resulting curve highlights the underlying shifts in momentum from the batting team's perspective.

1.3.3 Pressure Index (Benchmark for Comparison)

To validate our commentary-driven momentum index, we compare it against a simplified Pressure Index (PI) computed purely from scorecard variables. We adopt the variant of PI used in prior literature 1 & 2, omitting the wicket-weight term:

$$\text{PI}(t) = \left(\frac{\text{ReqRate}(t)}{\text{InitialRR}}\right) \times 100 + \left(\frac{\text{BallsLeft}(t)}{\text{TotalBalls}}\right)\left(\frac{\text{RunsLeft}(t)}{\text{Target}}\right), \tag{1.3}$$

where InitialRR = Target/TotalBalls, ReqRate(t) = (RunsLeft(t)/BallsLeft(t)) $\times$ 6, and the other quantities are defined in Section 1.2. We scale PI to the range $[0, 1]$ using min–max normalization to facilitate visual comparison with our momentum index.

1.3.4 Implementation Details

We implement the pipeline in Python using `pandas` for data handling, `nltk` for VADER sentiment analysis, and `scikit-learn` for scaling operations. All commentary lines without text or with missing match context are skipped, and the final momentum curve is indexed by the derived ball identifier (`Over_Ball_new`) to align with the innings timeline.

1.4 Results

1.4.1 South Africa vs Bangladesh (T20I, Potchefstroom, 2017-10-29)

Figure 1.1 compares the proposed commentary-driven momentum (rolling adjusted sentiment; orange) with a scaled Pressure Index (PI; blue) for the second innings of the match (Bangladesh chasing). Qualitatively, the momentum curve exhibits short-lived positive surges in the opening overs, followed by a decline toward neutral/negative values through the middle phase. In contrast, the scaled PI remains low in the powerplay and rises gradually, with a pronounced acceleration from approximately over 12 onward, indicating increasing chase difficulty.

These trajectories are consistent with the final result (South Africa won by 83 runs): as the chase stagnated in the middle and late overs, the commentary reflected mounting pressure (negative or near-neutral sentiment), while PI escalated sharply, signaling the growing gap between resources and requirement. Notably, the momentum curve captures subtle narrative cues (e.g., sequences of dot balls, "beaten"/"tight over" phrases, and near-chances) that do not immediately alter the scorecard but correspond to perceptible shifts in perceived control. The broad inverse relationship between momentum and PI in this innings supports our core hypothesis: commentary-derived sentiment provides a complementary, narrative-centric view of match flow.

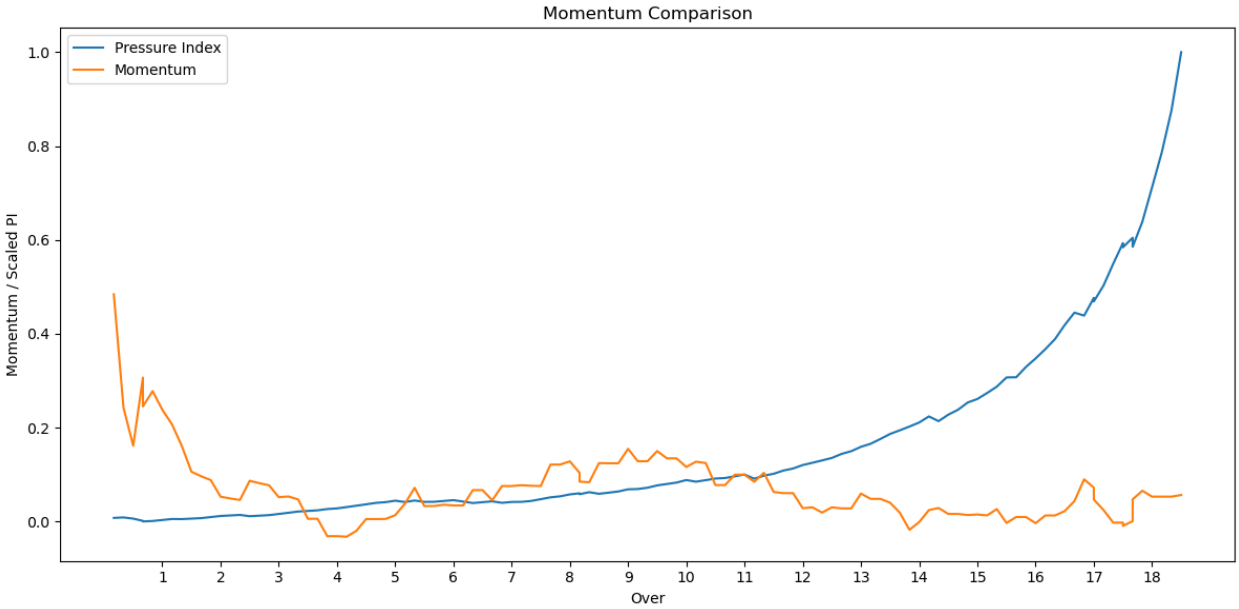

Figure 1.1: Momentum (rolling adjusted sentiment, orange) vs. scaled Pressure Index (blue) of Bangladesh for South Africa vs Bangladesh, T20I . Bangladesh batted second.

1.4.2 Afghanistan vs South Africa (ODI , Cardiff, 2019-06-15)

Figure 1.2 compares the proposed commentary-driven momentum (rolling adjusted sentiment; orange) with a scaled Pressure Index (PI; blue) for the second innings of the match (South Africa

chasing). The momentum curve begins slightly negative, reflecting early dot-ball pressure, before gradually stabilising around neutral values through the middle overs. The scaled PI, in contrast, starts high and steadily decreases across the innings, indicating that the chase became progressively easier as wickets were preserved and the required run rate dropped.

These patterns are consistent with the match outcome (South Africa won by 9 wickets). Once the initial pressure was absorbed, the commentary-derived momentum remained stable, with small positive fluctuations around overs 10–25 corresponding to partnerships building and boundaries being scored at regular intervals. Meanwhile, PI declined sharply in the final phase of the innings, capturing the near certainty of a successful chase. Notably, the commentary-based momentum captures narrative elements such as "steady accumulation", "good partnership" and "lofted for four", which signal increasing batting control even before the PI drops dramatically. This supports our hypothesis that commentary-derived sentiment provides early insight into shifts in match flow.

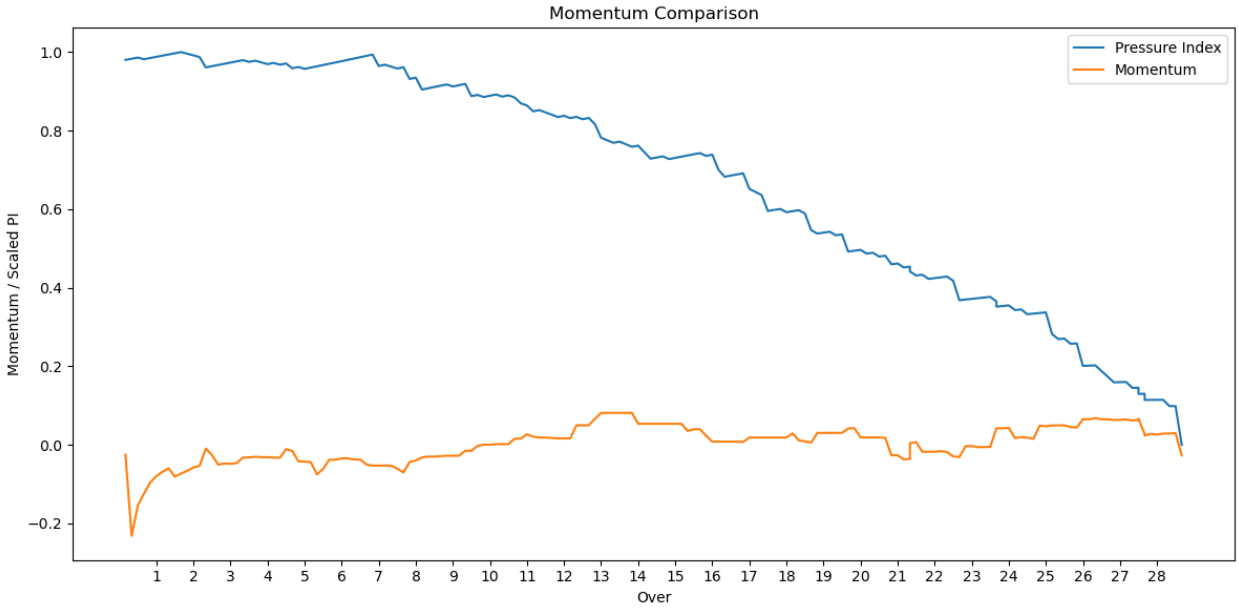

Figure 1.2: Momentum (rolling adjusted sentiment, orange) vs scaled Pressure Index (blue) of South Africa for Afghanistan vs South Africa, ODI. South Africa batted second.

1.4.3 England vs West Indies (T20I , Chester-le-Street, 2017-09-16)

Figure 1.3 compares the proposed commentary-driven momentum (rolling adjusted sentiment; orange) with a scaled Pressure Index (PI; blue) for the second innings of the match (England chasing). The momentum curve shows positive surges in the powerplay overs (1–6), reflecting boundary scoring and initial intent, before oscillating around neutral values during the middle overs. The scaled PI, on the other hand, remains low early and begins to climb gradually from around over 12, indicating rising chase difficulty as wickets fell and the required run rate tightened.

These observations align with the final result (West Indies won by 21 runs). After a promising start, England's momentum flattened and eventually declined in the latter half of the innings, while PI rose sharply in the closing overs. The commentary-derived momentum also captured narrative signals (e.g., "tight over", "beaten", "dot balls") that indicated mounting pressure, even when the required run rate appeared manageable. This match again demonstrates the complementary nature of commentary-based momentum: it highlights intangible factors of control and confidence that traditional scorecard metrics may miss.

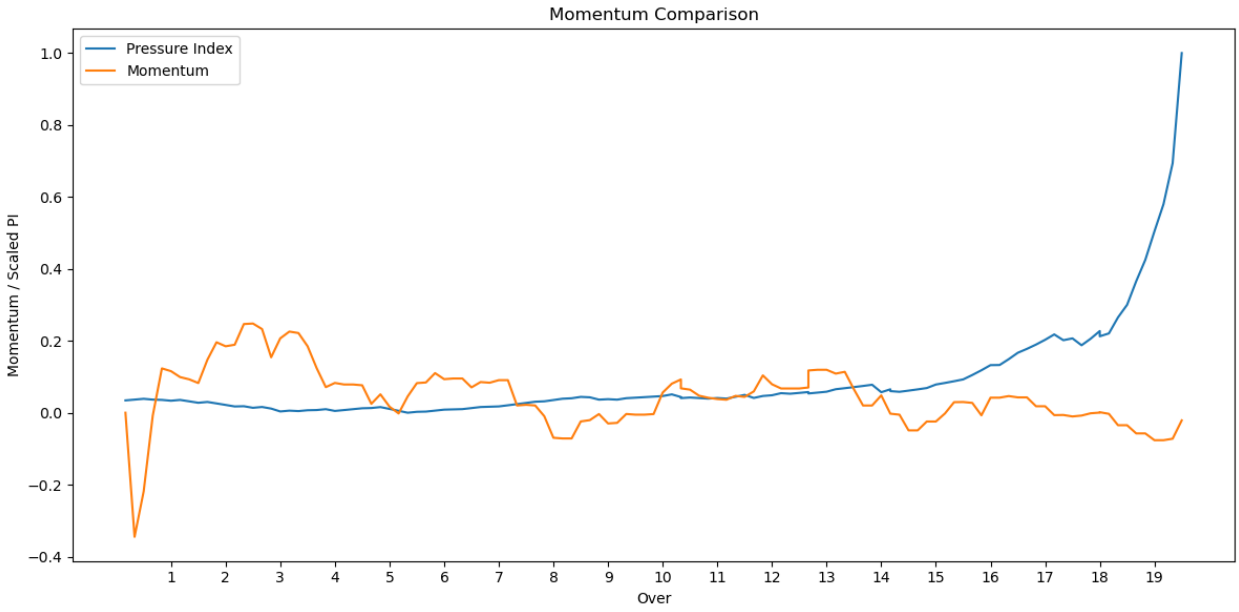

Figure 1.3: Momentum (rolling adjusted sentiment, orange) vs scaled Pressure Index (blue) of England for England vs West Indies, T20I . England batted second

1.5 Conclusion

Our experiments demonstrate that the commentary-derived momentum curve captures short-term fluctuations and narrative cues that PI often smooths out. PI tends to be a more aggregated and gradual indicator, driven primarily by changes in required run rate and remaining resources, whereas the momentum curve reacts immediately to events such as boundaries, wickets, or sequences of dot balls. For example, momentum spikes can be observed even during overs with no change to the scorecard (e.g., a dropped catch or a tight over with multiple appeals), offering early insight into shifts in perceived control.

An additional advantage of the commentary-driven approach is its general applicability. While PI can only be computed when complete structured data (target, runs left, balls left) is available, the momentum metric can be generated for any match or innings that has commentary data, including historical games or matches where only partial scorecard data is recorded. This flexibility makes it particularly useful for retrospective analysis or matches with limited data availability.

Overall, the commentary-driven momentum index complements traditional scorecard-based metrics by providing a narrative-centric, fine-grained view of match dynamics. In future work, richer natural language processing (NLP) techniques (e.g., transformer-based sentiment models or context-aware classification) could be applied to further improve the alignment of sentiment with match events and reduce the dependence on keyword lists. Integrating the momentum metric with PI and other structured features could also enable real-time predictive models for match outcomes.

References

1. Bhattacharjee, D., & Talukdar, P. (2019). Predicting outcome of matches using pressure index: evidence from Twenty20 cricket. Communications in Statistics - Simulation and Computation, 49(11), 3028–3040. https://doi.org/10.1080/03610918.2018.1532003

2. Shah, Parag & Shah, Mitesh. (2014). Pressure Index in Cricket. IOSR Journal of Sports and Physical Education. 1. 09-11. 10.9790/6737-0150911.