Research Article

Engineering: Open Access

Impact Calculation of The Players Using the Cricket Commentary Corpus

Aman Goel[1]* and Piyush Pratap Singh[2]

[1,2]School Of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India.

[*]**Corresponding Author**
Aman Goel, School of Computer and Systems Sciences, Jawaharlal Nehru University, India.

Submitted: 2023, June 26; **Accepted**: 2023, July 28; **Published**: 2023, Aug 04

Citation: Goel, A., Singh, P. P. (2023). Impact Calculation of The Players Using the Cricket Commentary Corpus. *Eng OA, 1*(2), 48-53.

Abstract

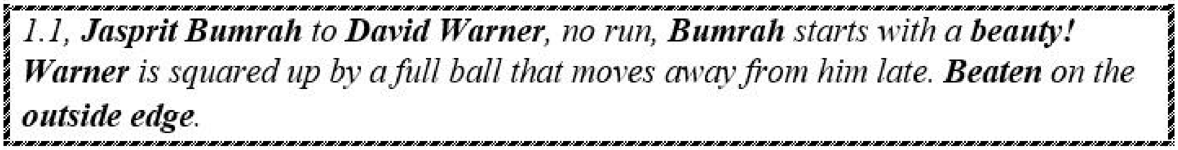
This research paper focuses on the use of natural language processing techniques for extracting insights from cricket commentary. This study proposes a framework that considers the commentary text, evaluates each player's performance, and gives value-based scores to define their impacts. This research relies on sentiment analysis, topic modeling, and NER to analyze a large corpus of cricket commentary data and develop a model for calculating the impacts of the players. The findings of this study will be of great interest to sports analysts, coaches, players, and fans of the game. This research explores cricket analytics and performance evaluation evolution, utilizing advanced textual analysis on cricket commentary to uncover hidden dimensions of player performance. This study has the potential to revolutionize player performance evaluation in cricket by identifying "hidden gems" that may have been overlooked based on conventional statistics alone.

Keywords: Cricket Commentary, Sentiment Analysis, NLP, Named Entity Recognition (NER), Impacts Score, MIP

1. Introduction

Cricket is a popular sport played in many coun- tries around the world, it is among the top-viewing sports in the world with an estimated fan base of nearly 2.6 billion1, with millions of fans tuning in to watch matches on television or streaming plat- forms. Furthermore, cricket commentary signifi- cantly impacts how the game is broadcast to the viewers, making the game appear very intriguing. The commentary during a cricket match provides a wealth of information on various aspects of the game, including player performance, strategy, and tactics. With the increasing popularity of data- driven approaches to sports analysis, researchers have begun to explore the use of natural lan- guage processing techniques to extract insights from cricket commentary.

This research takes the cricket commentary corpus of the Indian Premier League (IPL) as their primary source, which will play a pivotal role in calculating the impacts of the players. This paper proposes a framework that considers the commentary text, evaluates each player's per- formance, says something important about them, and gives value-based scores to play a significant role in defining their impacts. Computationally, the impact of cricket players is largely unknown.



*1.1, **Jasprit Bumrah** to **David Warner**, no run, **Bumrah** starts with a **beauty!** **Warner** is squared up by a full ball that moves away from him late. **Beaten** on the **outside edge**.*

Figure 1: Sample Commentary with Associated Aspects

Cricket statistics have been widely used to evalu- ate players. However, these metrics only provide a macroscopic view of a player's performance. Though they have a huge amount of data related to a player, they lag in capturing the fine-grained details associated with them. These problems can be conquered using the cricket text commentary, as this corpus provides real-time or finer details about each moment of the match.

A Commentary text can be divided into two parts: *fixed* and *variable*.

1. Fixed structure: Over [over number]. [ball num- ber]: [Bowler

[1]https://www.thecollector.com/history-of-cricket-worlds-second-most-popular-sport/

name] to [Batsman name], [out- come].
2. Variable structure: [Bowler's action/delivery], [Batsman's reaction/outcome of delivery].

Please refer to Figure 1. that provides us with an essential understanding of the variable struc- ture of the commentary, which includes associated aspects related to both the batsman and the bowler. Such terminologies in the commentary play a defining role in assessing the impact of cricket players.

This research will be relying on NLP tech- niques such as sentiment analysis, topic modeling, and named entity recognition that are well-suited to analyze the cricket commentary data, as they can help to identify key themes, emotions, and entities mentioned during a match.

This research paper will explore the applica-tion of these techniques to a large corpus of cricket commentary data, with the aim of developing a model for calculating the impact of the players. The analysis will focus on both batsmen and bowlers, fielding, and other aspects of the game. The findings of this research will be of great inter- est to sports analysts, coaches, and players, as well as fans of the game who are interested in under- standing the factors that contribute to a team's success or failure. In addition to this, this research also evaluates the term" Most *Impactful Player*" **(MIP)** in the course of this research (for a match or season).

The subsequent section presents a concise sum- mary of the existing research in this field and offers an overview of cricket studies, followed by comprehensive discussions on player impact calcu- lation using the cricket commentary corpus. Let's embark on an intriguing journey as we explore the impact calculation framework in depth. Through careful analysis of its results and conclusions, we will uncover hidden patterns and unravel the true essence of player performance.

2. Literature Review

The literature survey has been done in three dimensions, *evolution of Cricket Analytics and Performance Evaluation Methods, textual analy- sis, and analysis using commentary corpora.* This section reviews some prior work on particular scenarios, which are highlighted as follows:

2.1 Evolution of Cricket Analytics and Performance Evaluation Methods

In cricket analytics, F C Duckworth and A J Lewis introduced the D/L method for resetting tar- gets in rain-interrupted matches [1]. It ensures fair- ness by considering the shortening of the game and is easy to apply using a simple table of numbers and a calculator. The ICC adopted this method in 1998 for resetting targets in international cricket matches.

Researchers have explored optimal team selec-tion and batting order strategies in addition to target resetting. JM Norman and SR Clarke proposed a method using simulated annealing to determine optimal or nearly optimal batting orders in one-day cricket [2]. (John Norman and Stephen Clarke 2007) showed that variable bat- ting orders based on the state of the game increase expected scores across all cricket forms using dynamic programming. Integer program- ming models were employed by (Gerber and Sharp, Sharp et al [3]. Hermanus Hofmeyr Lemmer for optimal team selection in ODIs and T20s [4, 5].

Bharathan et al. proposed a methodology for player evaluation in team selection, considering multiple dimensions such as batting and bowl- ing, roles, context, and opponents [6]. They achieved an 83 percent accuracy in predicting team selec- tion using player evaluation utility. Other works include Silva and Swartz analyzing ODI statis- tics, showing no competitive advantage from win-ning the coin toss and increased winning proba-bility on home fields [7].

2.2 Textual Analysis

C.J. Hutto and Eric Gilbert wrote works describ- ing the process of sentence analysis using the parsimonious analyzer VADER, as documented in [8]. They used rule-based sentiment analy-sis. Researchers created a gold-standard sentiment lexicon for this model by combining qualitative and quantitative techniques. The lexicon was then empirically validated using particularly receptive microblog-type contexts. The grammatical and syntactic conventions of human speech are cap- tured by five generic rules, on that basis VADER combines significant lexical features. San Vicente and Saralegi outlined three techniques for cre- ating polarity lexicons: translation of pre-existing lexicons from other languages, extraction of polar- ity lexicons from corpora, and sentiment anno- tation [9]. Kusum Yadav et al [10]. have done their research on safe frameworks for feedback analysis. They described the source and the precise meth- ods involved in sentiment analysis and discussed how a statement might be awarded a valence- based rating. A little work has been done on Entity-Level Sentiment Analysis(ELSA) or Target sentiment analysis. Entity-level sentiment analysis forecasts sentiment regarding entities mentioned in a given text. Recently, ELSA was used by Xue- Yong Fu et al [11]. to examine English telephone conversation transcripts in contact centers in order to offer business insight. For this, they presented two methods: the CNN supplemented with some heuristic rules and the DistilBERT model. Similar to this, Jin Ding et al [12]. created and devel- oped a software called SentiSW, that essentially combines sentiment classification and entity recog- nition. SentiSW can categorize issue comments into ¡sentiment, entity¿ tuples. They used some open-source project comments from GitHub.

2.3 Analysis Using Commentary Corpus:

The sports news was created by Zhang et al [13]. using text commentary on ongoing events. In this study, they extracted the pertinent sentences from the documents using a machine learning approach. Using a "probabilistic sentence selec- tion" algorithm, news items are extracted from live football commentary while redundancy is minimized. Behera et al [14]. used short textual commentary to extract the players' strengths and weaknesses. They used biplots to plot a rela- tionship between bowler and batsman using the dimensionality reduction technique. Arif et al [15]. used text commentary to identify the strengths and weaknesses of cricket bowlers. An a priori algorithm is applied to a player's data by taking into account factors such as the home venue or away playing conditions.

A number of rules can be used to distinguish between different cricket events and commentary. The events mentioned can be linked to the match reports, as demonstrated by Mukta and Arefin [16].

It is also possible to find several research papers that aim to study the automatic genera- tion of live sports commentaries for video games, such as those by Zheng and Kudenko or media broadcasters Nijholt et al [17, 18]. The inher- ently valuable information in cricket, includes runs scored, performance indicators, batting averages, bowlers' given runs, and wickets per innings. The commentary text contains a lot of useful game information overall. Additionally, they provided a double banalization technique that divides the text into manageable chunks for the key subtitles [19]. Manish Gupta created the event-linking sys- tem known as CricketLinking, which accomplishes two tasks: it locates events mentioned in reports and then connects them to a group of balls.

After *analyzing* the state-of-the-art research mentioned above, it becomes clear that the objec- tive of the current study is different. None of the techniques outlined in the previous studies can be directly applied in the context of cricket text commentary. Limited research has been conducted on cricket commentary textual data, indicating a *research gap* in this relatively new and niche area with significant scope for exploration. There are very few works that are being done using the commentary text but none of the studies emphasize anything about any way or method of evaluating impactful players from the data, as this information can play a crucial role in various decision-making scenarios. Moreover, it can play a great impact *in enhancing the value of the play- ers who perform well in various stages of a match (irrespective of the winning or losing cause), but get de-emphasized due to certain scenarios like the team losing that match or some other players performed well in the later part of the game.*

3. Impact Calculation Framework

This study proposes a methodology for calculating the impact of cricket players on the game using commentary data. The proposed method- ology consists of four main parts: preprocessing, NER tagging, polarity checking and classification, and impact calculation. This paper describes each of these parts in detail and also provides a way to decide the impact of each player on the game irre- spective of whether their team wins or loses. This approach will significantly show the contribution of the players to the game.

3.1 Preprocessing

The first step in this methodology is to preprocess the commentary data., This involves removing any irrelevant text from the corpora. Basically, the phase deals with the refinement of the collection of textual data. Additionally, this part removes stop words and punctuation marks to ensure that only meaningful text is used in subsequent analysis. Finally, this part performs stemming and lemmatization to reduce the dimensionality of the data.

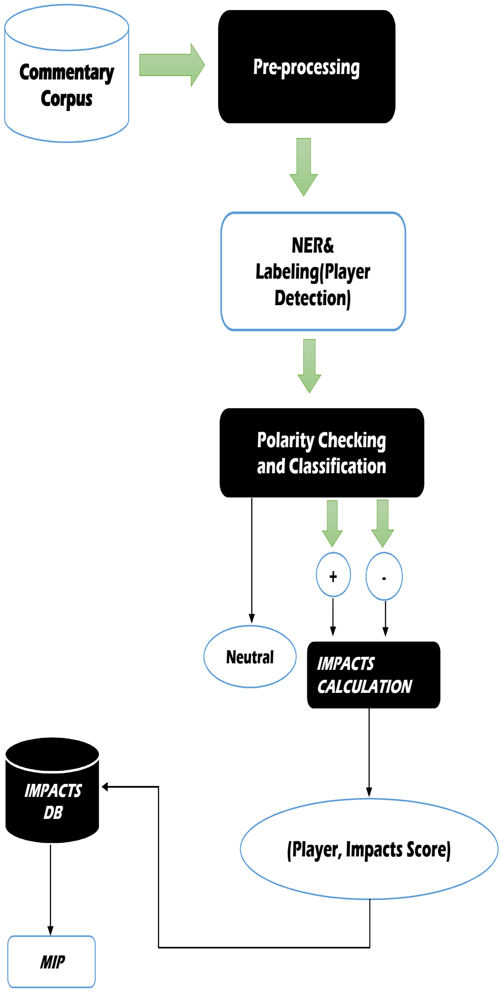


Figure 2: Proposed Approach Workflow

3.2 Named Entity Recognition (NER) Tagging

After the preprocessing stage, the next phase involves performing named entity recognition (NER) tagging on the data. This process is essen- tial for identifying the names of players men- tioned in the commentary. To accomplish this, we have employed a state-of-the-art, pre-trained NER model, such as the widely used Stanford NER or SpaCy NER, which are renowned for their accuracy and effectiveness in extracting named entities. The identified player names will be stored in a separate list for further analysis.

3.3 Polarity Checking and Classification

After identifying the player names, the method- ology will move to perform polarity checking and classification on the commentary text to identify the sentiment associated with each player's per- formance. n order to achieve this, a pre-trained sentiment analysis model will be utilized to clas- sify the commentary text into positive, negative, or neutral categories. The neutral text will be filtered out, as it is not expected to contribute significantly to determining player impact. Addi- tionally, the model will assign a **polarity score** to each player's performance to indicate the degree of positivity or negativity in the commentary asso- ciated with their performance. For example, if a player's performance is associated with a high polarity score, it indicates that their performance was highly appreciated by the commentators.

3.4 Impact Calculation

The final step of this methodology is to calculate the impact of each player using the NER-tagged data and the polarity scores. To determine the impact of each player on the game, irrespective of winning or losing, this work proposes the following set of rules:

1. For each player, the calculation of their impact score by summing up the polarity scores associ- ated with their performance in the commentary data will be done.
2. After that the normalization of the impact scores of all players will be taken into account for the number of overs played by their team in the match.
3. Finally, it will calculate the relative impact of each player by comparing their normal- ized impact score with the average normalized impact score of all players in the match.

We can do this task by incorporating the following terminologies: For a batsman, the performance metrics will be calculated based on the runs scored on the ball, the number of balls faced by the batsman, the number of boundaries hit by the batsman on the ball, and the match situation at the ball, such as power-play, death overs, or difficult parts of the match.

Similarly, for a bowler, the following perfor- mance metrics will be calculated for each ball bowled: wickets taken on that ball, number of dot balls bowled by the bowler, number of boundaries conceded by the bowler, and the match situation at the ball.

Above, terminology can be simplified or evalu-ated mathematically using the following methods:

For a **Batsman**, this can be calculated using a formula such as:

Performance Score = (Runs Scored / Balls Faced) * (Boundaries Hit / Balls Faced) * (Match Situation Factor)

For a **Bowler**, this can be calculated using a formula such as:

Performance Score = (Wickets Taken / Balls Bowled) * (Dot Balls Bowled / Balls Bowled) * (Match Situation Factor)

Here, ***Match Situation Factor*** is a metric that takes into account the match situation at the time the player batted or bowled. It can be assigned based on factors such as the powerplay, the death overs, the type of pitch, the tough phase of the match, etc.

Calculation of the ***Impact Score*** of the player for the match can be carried away using the formula:

Impact Score = *(Performance Score) × [Polarity Score(positive) - Polarity Score(negative)] × (Difficulty Score)*

The ***Difficulty Score*** is used to account for the difficulty of the match situation that the player was performing in. This is important because *a player's impact in a match depends not only on their own performance but also on the quality of the opposition team and the match conditions.* For example, a player who performs well against a strong opposition team in challenging match con- ditions (such as a difficult pitch or a tight chase) is likely to have a greater impact on the match than a player who performs well against a weaker team in easier match conditions.

By incorporating a Difficulty Score into the impact calculation, we can give more weight to the performance of players who perform well in challenging match situations, and adjust the impact score accordingly.

The proposed methodology provides a fair and unbiased approach for evaluating the impact of each player on a cricket match, irrespective of the overall match outcome. This approach is important as it can help to recognize the individual contributions of the players, which may not always be apparent from the overall match result. The methodology achieves this by incorporating factors such as sentiment analysis and difficulty score, which provide a more comprehensive evaluation of player performance.

However, it is important to note that there may be potential limitations associated with this methodology, such as the quality and reliability of the commentary data, the accuracy of the sentiment analysis, and the appropriateness of the performance metrics and weightings used. Further research could explore ways to address these limitations and improve the methodology.

3.5 Concluding statement for Proposed Work

The presented methodology offers an unbiased assessment of player impact in cricket matches, considering factors like sentiment analysis and difficulty scores. However, potential limitations include data quality and analysis accuracy, which could be addressed in future research. Overall, the proposed framework provides valuable insights for player evaluation and strategic decision-making. By identifying the players who had the highest impact on the match outcome.

4. Results and Discussion

Referring to table 1. that shows an example related to this approach's result, the Impacts Score of a few players from the ICC World Cup 2011 final are highlighted. Overall, the impact calculation using the cricket commentary corpus has proved to be a useful tool in analyzing the performance of players in different match situations. The results can provide valuable insights for team selection and strategy planning in future matches.

Players Name	Impacts Score
Gautam Gambhir (India)	0.594
MS Dhoni (India)	0.536
Mahela Jayawardene (Sri Lanka)	0.967
Virat Kohli (India)	0.392
Kumar Sangakkara (Sri Lanka)	0.252
Yuvraj Singh (India)	0.298
Lasith Malinga (Sri Lanka)	0.143

Table 1: Impacts Table

Please refer to Fig 3. to visualise the table using a bar chart.

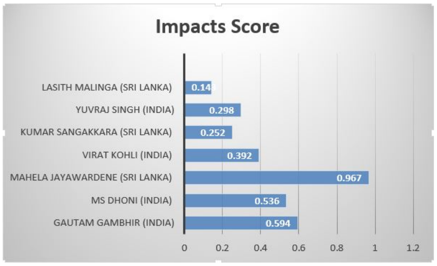


Figure 3: Visualisation of the result

5. Conclusion

This research proposes a framework for calculating the impacts of cricket players by analyzing cricket text commentary using NLP techniques such as sentiment analysis, topic modeling, and named entity recognition. The proposed frame- work considers the commentary text, evaluates each player's performance, says something impor- tant about them, and gives value-based scores to significantly define their impacts. The literature review is conducted on two dimensions: textual analysis and analysis using commentary corpus.

The proposed framework provides a model for cal-culating the impact of players, with a focus on both batsmen and bowlers, fielding, and other aspects of the game. The findings of this research will be useful to sports analysts, coaches, and players, as well as fans of the game who are interested in understanding the factors that con-tribute to a team's success or failure. Also, it is a great decision-maker for the players who get de-emphasized due to certain scenarios as mentioned in this paper. The current research can be extended to various dimensions, such as Multi-Lingual Analysis, and Real-time Analysis, Integrating with external data sources can provide context for analyzing player performance, includ- ing correlations with factors such as injuries or team dynamics.

6. Declarations

6.1 Funding

Not applicable.

6.2 Author's Contribution

We, the authors of this research paper, collectively declare significant contributions to the study. The specific contributions are as follows:

- *Aman Goel* played a crucial role in this research study. He contributed significantly to the study's conceptualization, data collection, and preprocessing. Aman implemented natural language processing techniques, developed a framework for player performance evaluation, conducted statistical analysis, and actively participated in manuscript drafting and revision. His contributions shaped the study's direction and enriched its outcomes.
- *Dr. Piyush Pratap Singh,* as the co-author and supervisor, made significant contributions throughout the research study. He provided guidance in conceptualization, methodology, and data analysis. His expertise in natural language processing and cricket commentary enhanced the research findings. Dr. Singh offered critical feedback and support during the writing and ensured the scholarly rigor of the manuscript.

6.3 Conflict of Interest

Not applicable

6.4 Data Availability Statement

The data used in this research paper is available upon request from the authors and is also publicly accessible online.

References

1. Duckworth, F. C., & Lewis, A. J. (1998). A fair method for resetting the target in interrupted one-day cricket matches. Journal of the Operational Research Society, 49(3), 220-227.
2. Norman, J. M., & Clarke, S. R. (2010). Optimal batting orders in cricket. Journal of the Operational Research Society, 61(6), 980-986.
3. Gerber, H., & Sharp, G. D. (2006). Selecting a limited overs cricket squad using an integer programming model. South African Journal for Research in Sport, Physical Education and Recreation, 28(2), 81-90.
4. Sharp, G. D., Brettenny, W. J., Gonsalves, J. W., Lourens, M., & Stretch, R. A. (2011). Integer optimisation for the selection of a Twenty20 cricket team. Journal of the Operational Research Society, 62(9), 1688-1694.
5. Lemmer, H. H. (2013). Team selection after a short cricket series. European Journal of Sport Science, 13(2), 200-206.
6. Bharathan, S., Sundarraj, R. P., Abhijeet, S., & Ramakrishnan, S. (2015). A self-adapting intelligent optimized analytical model for team selection using player performance utility in cricket. In 9th MIT Sloan Sports Analytics Conference, MIT, Boston (pp. 1-11).
7. De Silva, B. M., & Swartz, T. B. (1998). Winning the coin toss and the home team advantage in one-day international cricket matches. Department of Statistics and Operations Research, Royal Melbourne Institute of Technology.
8. Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 216-225).
9. San Vicente, I., & Saralegi, X. (2016, May). Polarity lexicon building: to what extent is the manual effort worth?. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 938-942).
10. Yadav, K., Pandey, M., & Rautaray, S. S. (2017, February). A proposed framework for feedback analysis system using big data tools. In 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) (pp. 542-545). IEEE.
11. Fu, X. Y., Chen, C., Laskar, M. T. R., Gardiner, S., Hiranandani, P., & Tn, S. B. (2022). Entity-level sentiment analysis in contact center telephone conversations. arXiv preprint arXiv:2210.13401.

12. Ding, J., Sun, H., Wang, X., & Liu, X. (2018, June). Entity-level sentiment analysis of issue comments. In Proceedings of the 3rd International Workshop on Emotion Awareness in Software Engineering (pp. 7-13).
13. Zhang, J., Yao, J. G., & Wan, X. (2016, August). Towards constructing sports news from live text commentary. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1361-1371).
14. Behera, S. R., Agrawal, P., Awekar, A., & Vedula, V. S. (2019, December). Mining strengths and weaknesses of cricket players using short text commentary. In 2019 18th IEEE international conference on machine learning and applications (ICMLA) (pp. 673-679). IEEE. Available from: https://doi.org/10. 1109/ICMLA.2019.00122.
15. Arif, S., Umair, M., Naqvi, S. M. K., Ikram, A., & Ikram, A. (2018, June). Detection of bowler's strong and weak area in cricket through commentary. In Proceedings of the 2nd international conference on future networks and distributed systems (pp. 1-14).
16. Mukta, R. B. M., & Arefin, M. S. (2019). An Agent Based Parallel and Secure Framework to Collect Feedbacks. J. Comput., 14(6), 404-425.
17. Zheng, M., & Kudenko, D. (2010). Automated event recognition for football commentary generation. International Journal of Gaming and Computer-Mediated Simulations (IJGCMS), 2(4), 67-84.
18. Cheok, A. D., Romão, T., Nijholt, A., & Yu, G. (2014). Entertaining the whole world. In Entertaining the Whole World (pp. 1-3). London: Springer London.
19. Gupta, M. (2015, August). CricketLinking: linking event mentions from cricket match reports to ball entities in commentaries. In Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval (pp. 1033-1034).