Paper 136-2012

Correlating the Analysis of Opinionated Texts Using SAS® Text Analytics with Application of Sabermetrics to Cricket Statistics

Praveen Lakkaraju and Saratendu Sethi, SAS Institute Inc., Cambridge, MA

ABSTRACT

Cricket is a game similar to baseball. It is rich in statistics, and there are plenty of online discussions about the game and the players. Sabermetrics deals with using statistical methods to analyze baseball records. This paper presents the results of experiments with applying Sabermetrics-style principles to the game of cricket and correlating the findings with analysis of opinionated text. Examples use various products from the SAS® Text Analytics suite to demonstrate how structured and unstructured data analysis can be applied together to gain insights into real-world data.

INTRODUCTION

Cricket is a hugely popular game in India and some of the Commonwealth countries. Similar to baseball, it has a batter (called a batsman) and a pitcher (called a bowler), and the team that scores the higher number of runs wins the match. The goal of the batsmen is to score as many runs as they can; the goal of the bowlers is to get as many batsmen out as they can while also restricting the number of runs scored. However, the rules, grounds, and equipment are completely different. There are three different formats of the game at the international level: Test matches, One-Day International (ODI) matches, and Twenty20 International (T20I) matches. Although the basic rules of the game remain the same among these formats, the rules that pertain to the duration of the game and the result are different in each format.

Sabermetrics deals with the use of statistical methods to study baseball records. The people who practice sabermetrics are called sabermetricians. They frequently challenge the traditional metrics (such as batting averages) for measuring a player's performance, introducing new metrics to measure the true value of a player to a given team. Because cricket is similar to baseball, some of the fundamental ideas behind sabermetrics can also be applied to cricket. A growing number of cricket enthusiasts and statisticians around the world are researching ways to apply sabermetrics-style principles to cricket statistics (Borooah and Mangan 2010). With the advent of Twenty20, which is the shortest form of the game (each game lasts about 3.5 hours), a lot of leagues that include teams with private owners have been started in recent years. This has led to a tremendous rise in player salaries. The teams and owners are starting to look at a variety of metrics to make sure they invest in the most valuable players. Because it is new, the Twenty20 format does not yet have a sufficient amount of data to use for statistical analysis. So this paper focuses on the matches that belong to the ODI format. The basic principles can easily be applied to the Twenty20 format when a sufficient amount of data becomes available.

The ODI format of the game was introduced in the early 1970s. Since then about 4,000 matches have taken place at the international level. During the '70s, the '80s, and most of the '90s, the discussions about the game appeared predominantly in print media. As the Internet became popular in India and other developing countries during the late '90s and early 2000s, the discussions started shifting online. Fans express their opinions about the players and the teams in a variety of online forums. There is a lot of value in analyzing these unstructured data to understand how the fans feel about the players and teams.

This paper discusses the results of some experiments that analyze textual data about a few top players of recent years. Two performance metrics are developed to measure the batting performance of a given player, and then the performance of some of the players is correlated with the analysis of textual data that are related to them.

THE DATA

Cricinfo (ESPN 2012) is the source of all the scorecard and textual data that are used for the experiments in this paper. Cricinfo has been the most popular site for cricket fans around the globe for the past decade. SAS® Information Retrieval Studio, a component of SAS® Web Crawler, was used to collect all the scorecards for the ODIs played by the top 35 batsmen (chosen by the total number of runs scored) in that format from 2006 to 2011. SAS Web Crawler was also configured to download the articles from the site that mentioned these players and their corresponding comments. Custom document processers were created in SAS Information Retrieval Studio to parse the scorecards and articles and to extract the information that was needed to compute the performance metrics to be used in the experiments. Because this paper focuses on ODIs, another custom document processor that uses SAS® Content Categorization to filter out articles that were not related to ODIs was also created. All this processing resulted in scorecard information for about 1,000 ODIs, about 4,500 articles, and about 53,000 comments, which were all stored in XML format so that they could be easily processed using other SAS Text Analytics tools.

THE METRICS

As opposed to the traditional metrics that measure batting performance (such as the total number of runs, batting average, strike rate, and so on), this paper uses two new metrics: consistency-adjusted average (CAA) and batting impact score (BI). The following sections explain the methodologies for computing these metrics.

CONSISTENCY-ADJUSTED AVERAGE (CAA)

Borooah and Mangan (2010) used consistency-adjusted average (CAA) to (re)rank the top batsmen of all time in Test cricket. This paper uses the metric in the context of one-day internationals. As discussed in Borooah and Mangan, just looking at the average of a batsman over a period of time might not provide the full picture of how the batsman has performed. Averages can be inflated because of a few high scores. It is important to measure how consistent the batsman was in his performance. To measure consistency, the Gini coefficient is used (Gini 1912). The Gini coefficient measures the inequality among values of a frequency distribution. The value can range from 0 to 1. A low value indicates a more equal distribution (with 0 corresponding to complete equality), and a high value indicates a more unequal distribution (with 1 corresponding to complete inequality). In the context of one-day international cricket, the Gini coefficient is computed as

$$G = \frac{1}{2N^2\,\mu}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|R_i - R_j\right|$$

where *G* is the Gini coefficient, *N* is the number of completed innings, μ is the average score, and R_i is the number of runs scored by the batsman in the *i*th innings.

Then the consistency-adjusted average (CAA) is computed as

$$CAA = \mu * (1 - G)$$

where μ is the consistency-unadjusted average (that is, the regular average).

BATTING IMPACT (BI)

Batting impact (BI) score measures a player's performance in the context of a given match. It takes into account not only how many runs a player has scored but also the pace at which he scored the runs and the conditions under which he scored the runs. This is similar to some of the metrics discussed on the Impact Index website (Impact Index Cricket Pvt. Ltd. 2012). A BI score is assigned to every player who batted in a given match based on the following aspects of his performance:

- Runs Impact Score (RIS): This metric measures the ratio of the runs scored by the player against the mean runs for all players for the match. The RIS for a player *p* in match *m* is computed as

$$RIS_{pm} = \frac{Runs_{pm}}{BaseRuns_m}$$

$$BaseRuns_m = {\sum_{i=1}^{N} Runs_{im}}/{N}$$

  where $Runs_{pm}$ represents the runs scored by player *p* in match *m*, and *N* is the total number of players who batted in the match.

- Strike Rate Impact Score (SRIS): This metric assigns a positive score to the player if his strike rate is above the mean strike rate for the match and a negative score if it is below. If his strike rate equals the mean strike rate of the match, then his SRIS is equal to zero. The SRIS for a player *p* in match *m* is computed as

$$SRIS_{pm} = \frac{SR_{pm}}{{\sum_{i=1}^{N} Runs_{im}}/{\sum_{i=1}^{N} Balls_{im}}} - 1$$

  where SR_{pm} is the strike rate of player *p* in match *m*, and $Balls_{im}$ is the number of balls faced by player *i* in match *m*.

- Pressure Impact Score (PrIS): The score for this metric is a product of two variables: the situation in which the player comes in to bat and the player's RIS. The first variable is called the Pressure Factor (PF), which depends on the difficulty of the situation when the batsman comes to bat. A situation's difficulty is measured in terms of the number of wickets that have fallen and the mean runs for the match. The PrIS for a player *p* in match *m* is computed as

$$PrIS_{pm} = PF_{pm} * RIS_{pm}$$

$$PF_{pm} = \frac{(INW_{pm} * BaseRuns_m) - INS_{pm}}{BaseRuns_m}$$

where INW_{pm} is the number of wickets that have fallen in the innings when player *p* comes to bat in match *m*, and INS_{pm} is the score of the team when player *p* comes to bat in match *m*.

If player *p* is an opening batsman, then INW_{pm} is always 0. To account for this, all opening batsmen are given a PF of 0.5. In other words, he is expected to score at least half the number of mean runs for the match.

- Chasing Impact Score (ChIS): This special score is assigned to a player for staying not out in the second innings of a successful chase. If a player satisfies this criterion, then his ChIS is equal to his RIS; otherwise, it is equal to 0.

Based on the preceding four scores, the overall batting impact score (BI) of a player *p* in a match *m* is computed as

$$BI_{pm} = \frac{RIS_{pm} + SRIS_{pm} + PrIS_{pm} + ChIS_{pm}}{MaxBI_m}$$

where $MaxBI_m$is the highest batting impact score for match *m* and is used as a normalization factor.

SENTIMENT

A rules-based sentiment model was developed using SAS® Sentiment Analysis Studio to extract sentiment from the textual data that were collected about cricket players. Figure 1 shows a simple taxonomy with the names of the players as the top-level nodes. SAS Sentiment Analysis provides all the tools needed to develop statistical and natural language rules-based models for sentiment analysis (Albright and Lakkaraju 2010; Lange and Sethi 2011). You can use a combination of language dictionaries, linguistic constructs such as parts of speech, noun phrases, and co-reference resolution along with a range of operators (see Figures 1 and 2).

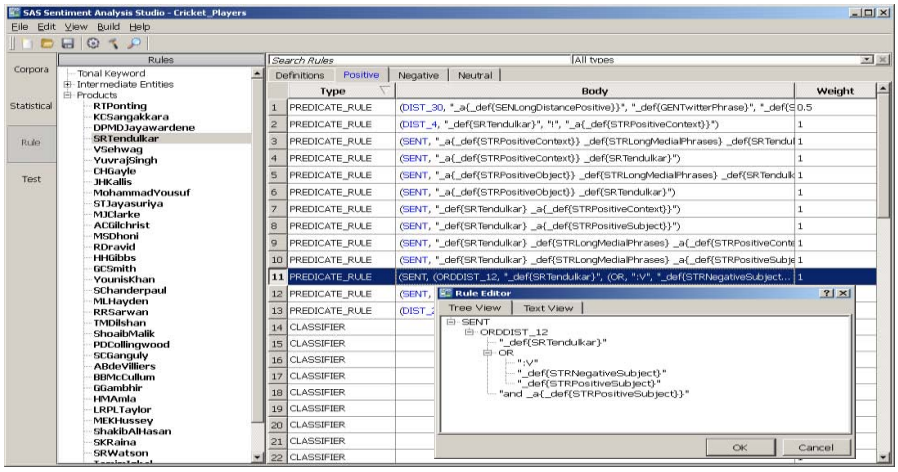

Figure 1. Taxonomy for Cricket Players in SAS Sentiment Analysis Studio

For the model used in this paper, some prebuilt dictionaries of polarity expressions were used in combination with other structural entities (Figure 2). Developing the rules for players' definitions was straightforward because there were only a handful of variations of their names. For example, for Virender Sehwag, variations of his name such as Veeru and Sehwag, were added as definitions. Rules for pronoun resolution were also added.

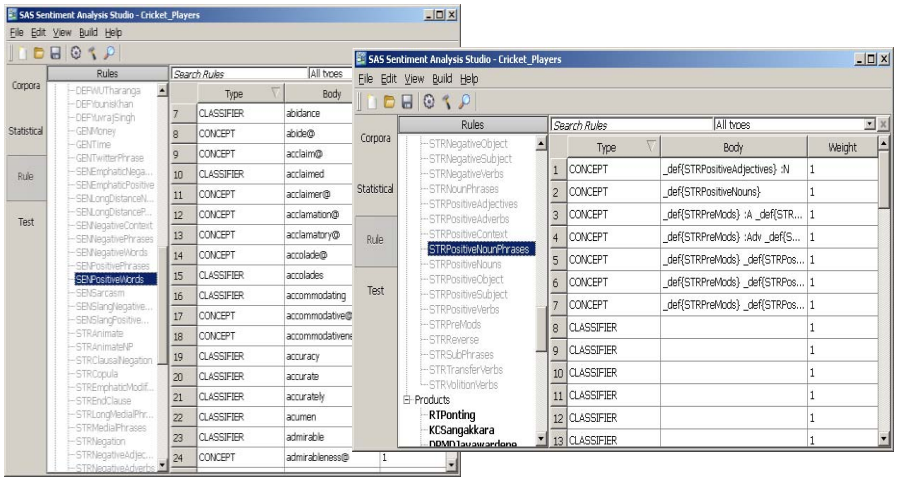

Figure 2. Example of a Dictionary of Positive Keywords and Noun Phrases

When the entities and definition rules were in place, some Boolean rules were created to extract sentiment for the players (Figure 1). A subset of documents was imported into SAS Sentiment Analysis Studio and tested to make sure the sentiment was being assigned correctly; in cases where it was not, the rules were updated accordingly (Figure 4).

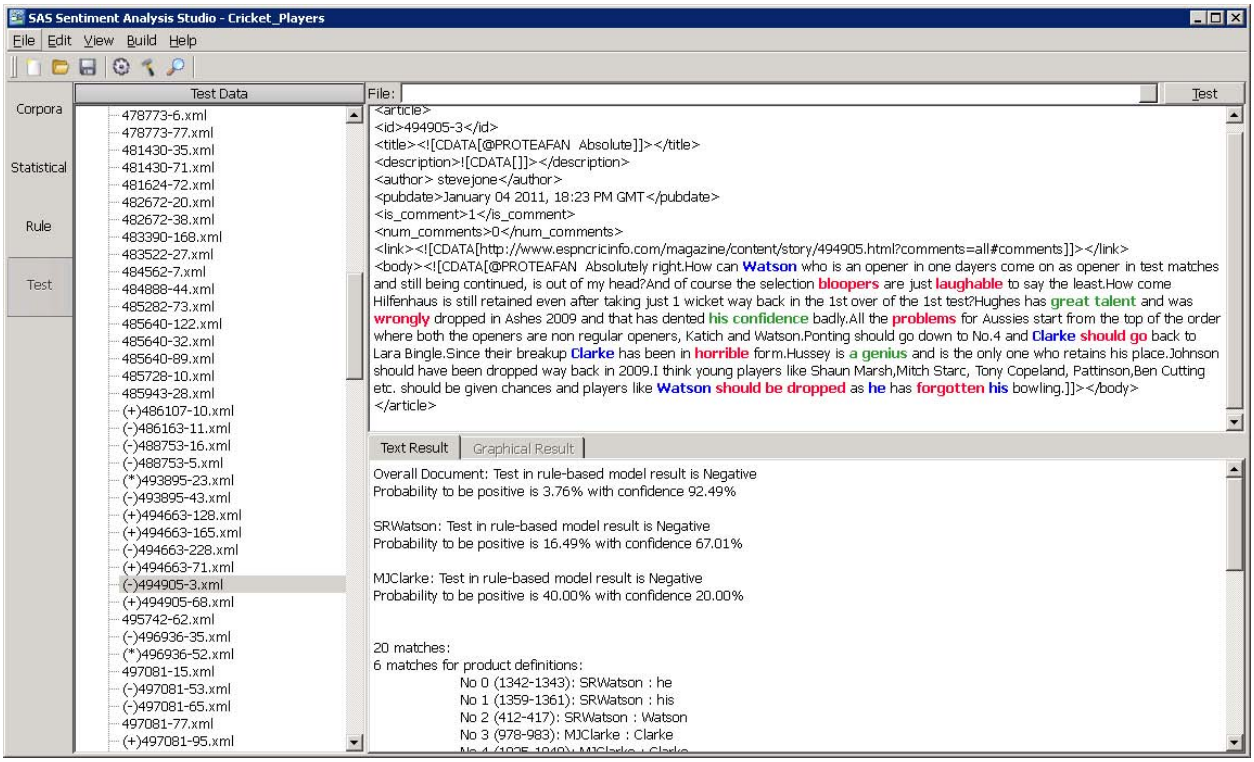

Figure 3. Example of Testing a Document in SAS Sentiment Analysis Studio

The model was then deployed in SAS® Sentiment Analysis Server and the full set of documents was scored for sentiment.

RESULTS

The two new metrics were applied to the top 25 batsmen in terms of the total number of runs scored by them between 2006 and 2011, and the batsmen were ranked again based on these metrics. Table 1 shows how the rankings of the top batsmen changed when they were ranked again by their consistency-adjusted average (CAA).

Rank by CAA	Player Name	Average[1]	Consistency-Adj Average (CAA)	Rank by Average	Diff in Rank
1	MS_Dhoni	51.18	35.36	2	1
2	S_Chanderpaul	52.12	33.83	1	–1
3	MEK_Hussey	47.09	32.33	6	3
4	AB_de_Villiers	50.66	30.97	3	–1
5	MJ_Clarke	45.94	29.58	7	2
6	JH_Kallis	47.66	28.67	4	–2
7	SR_Watson	45.52	26.89	8	1
8	Yuvraj_Singh	42.96	25.50	10	2
9	SR_Tendulkar	47.52	24.75	5	–4
10	KC_Sangakkara	41.17	24.23	12	2
11	G_Gambhir	43.09	23.80	9	–2
12	RT_Ponting	42.44	23.53	11	–1
13	PD_Collingwood	37.25	22.06	18	5
14	TM_Dilshan	37.55	20.75	17	3
15	V_Sehwag	41.11	20.69	13	–2
16	LRPL_Taylor	37.83	20.67	14	–2
17	GC_Smith	37.57	20.36	16	–1
18	CH_Gayle	37.64	19.59	15	–3
19	Shakib_Al_Hasan	33.59	19.46	22	3
20	DPMD_Jayawardene	36.44	19.39	19	–1
21	WU_Tharanga	36.06	17.96	20	–1
22	Younis_Khan	32.63	17.00	23	1
23	BB_McCullum	31.79	16.48	24	1
24	ST_Jayasuriya	34.53	15.67	21	–3
25	Tamim_Iqbal	28.44	13.34	25	0

Table 1. Top 25 Batsmen in ODIs Ranked by CAA: 2006–2011

[1]The averages in the table were computed based on the data that were collected. The actual averages might be slightly different.

Table 1 shows that when the top players were ranked again based on their CAA, Collingwood had the biggest jump in his ranking. Although he had been a consistent performer for England in the last few years, he has often been criticized for being unable to convert his starts. The rankings in Table 1 support that general opinion about him. The ranking for Tendulkar dropped by four points; over the last six years, Tendulkar had mixed performance in ODIs. He achieved one of the highest scores in ODIs (200*), which was also his personal best in 2010. However, he was very selective about participating in this format of the game and could not play for some period of time because of injuries. All this led to some inconsistency in his performance, and the rankings in Table 1 also show this inconsistency.

Table 2 shows the ranking of the top 25 batsmen based on their consistency-adjusted average batting impact (BI) score.

Rank by CABI	Player Name	Average[1]	Consistency-Adj Average BI	Rank by Average	Diff in Ranks
1	SR_Watson	45.52	0.2423	8	7
2	V_Sehwag	41.11	0.2030	13	11
3	S_Chanderpaul	52.12	0.1949	1	–2
4	MEK_Hussey	47.09	0.1948	6	2
5	SR_Tendulkar	47.52	0.1939	5	0
6	CH_Gayle	37.64	0.1868	15	9
7	AB_de_Villiers	50.66	0.1855	3	–4
8	JH_Kallis	47.66	0.1784	4	–4
9	MS_Dhoni	51.18	0.1730	2	–7
10	KC_Sangakkara	41.17	0.1644	12	2
11	GC_Smith	37.57	0.1623	16	5
12	ST_Jayasuriya	34.53	0.1597	21	9
13	TM_Dilshan	37.55	0.1593	17	4
14	WU_Tharanga	36.06	0.1579	20	6
15	G_Gambhir	43.09	0.1571	9	–6
16	MJ_Clarke	45.94	0.1554	7	–9
17	RT_Ponting	42.44	0.1553	11	–6
18	Yuvraj_Singh	42.96	0.1529	10	–8
19	BB_McCullum	31.79	0.1475	24	5
20	PD_Collingwood	37.25	0.1437	18	–2
21	Shakib_Al_Hasan	33.59	0.1334	22	1
22	Younis_Khan	32.63	0.1313	23	1
23	Tamim_Iqbal	28.44	0.1299	25	2
24	LRPL_Taylor	37.83	0.1266	14	–10
25	DPMD_Jayawardene	36.44	0.1253	19	–6

Table 2. Top 25 Batsmen in ODIs Ranked by CABI: 2006–2011

[1]The averages in the table were computed based on the data that were collected. The actual averages might be slightly different.

In Table 2, you can see that Sehwag, Jayasuriya, and Gayle had the biggest jumps in their rankings. All three of them are proven match winners for their respective teams. They have the reputation of being high-impact players, and the ranking in Table 2 supports that reputation.

Table 3 shows the top batsmen ranked by the number of documents that mentioned them over the last six years. The term “document” refers both to articles and to their comments. Each comment is considered to be a separate document.

Rank by Mentions	Player Name	Average[1]	Number of Mentions	Rank by Average	Diff in Ranks
1	SR_Tendulkar	47.52	7576	5	4
2	RT_Ponting	42.44	4420	11	9
3	V_Sehwag	41.11	3864	13	10
4	MS_Dhoni	51.18	3010	2	–2

5	MJ_Clarke	45.94	2189	7	2
6	MEK_Hussey	47.09	1760	6	0
7	GC_Smith	37.57	1668	16	9
8	Yuvraj_Singh	42.96	1613	10	2
9	G_Gambhir	43.09	1495	9	0
10	CH_Gayle	37.64	1487	15	5
11	SR_Watson	45.52	1345	8	–3
12	ST_Jayasuriya	34.53	1301	21	9
13	DPMD_Jayawardene	36.44	1288	19	6
14	JH_Kallis	47.66	1243	4	–10
15	TM_Dilshan	37.55	820	17	2
16	LRPL_Taylor	37.83	755	14	–2
17	PD_Collingwood	37.25	668	18	1
18	KC_Sangakkara	41.17	613	12	–6
19	S_Chanderpaul	52.12	533	1	–18
20	WU_Tharanga	36.06	379	20	0
21	BB_McCullum	31.79	291	24	3
22	AB_de_Villiers	50.66	265	3	–19
23	Tamim_Iqbal	28.44	178	25	2
24	Shakib_Al_Hasan	33.59	139	22	–2
25	Younis_Khan	32.63	111	23	–2

Table 3. Top 25 Batsmen in ODIs Ranked by Number of Mentions: 2006–2011

[1]The averages in the table were computed based on the data that were collected. The actual averages might be slightly different.

In Table 3, you can see that Tendulkar and Ponting, who are widely considered the two most popular and most successful ODI players of all time, rank at the top of the list.

The next set of reports takes the top six players between 2006 and 2011 in terms of the number of their mentions and correlates their sentiment analysis results with their performance metrics.

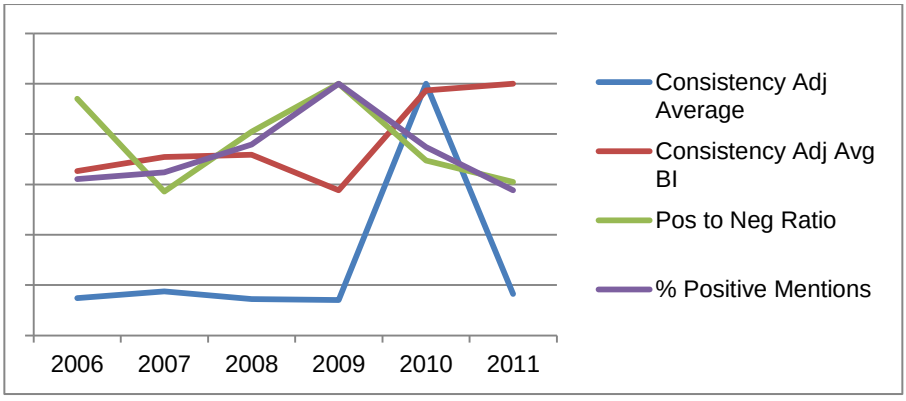

Figure 4. Tendulkar's Performance versus Sentiment

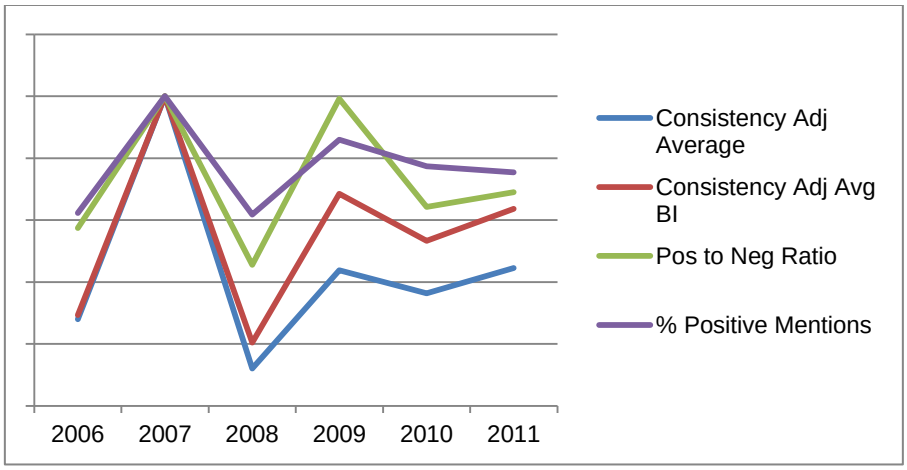

Figure 5. Ponting's Performance versus Sentiment

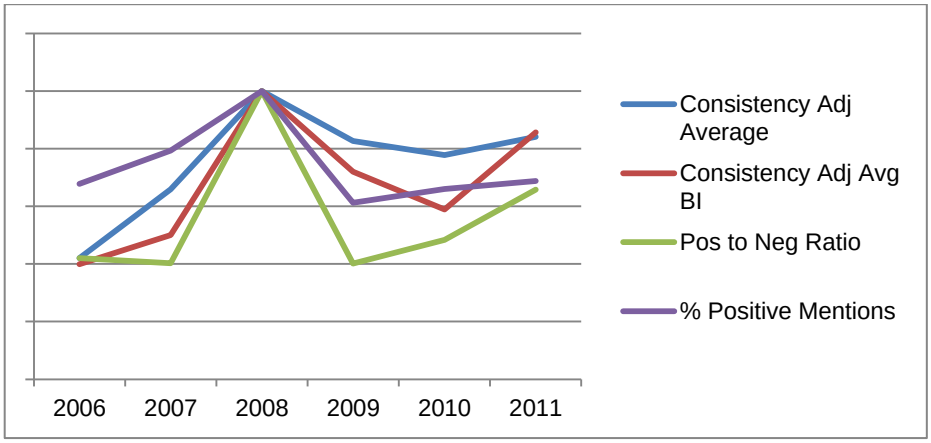

Figure 6. Sehwag's Performance versus Sentiment

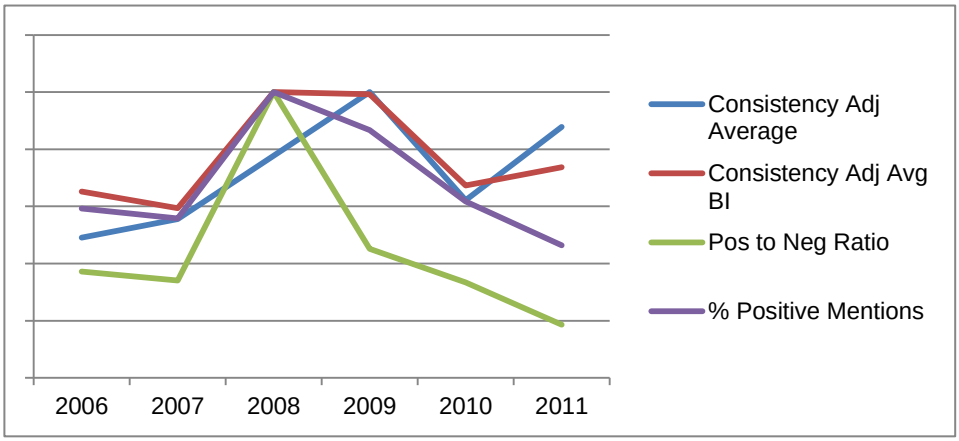

Figure 7. Dhoni's Performance versus Sentiment

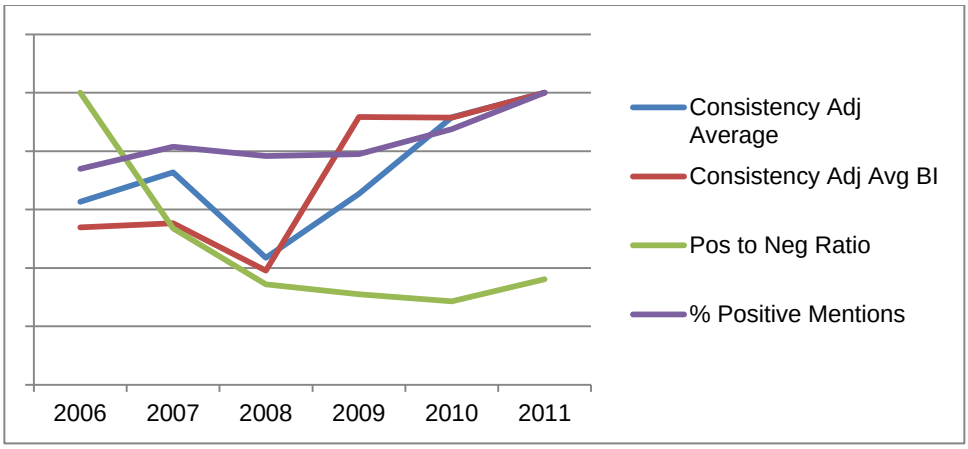

Figure 8. Clarke's Performance versus Sentiment

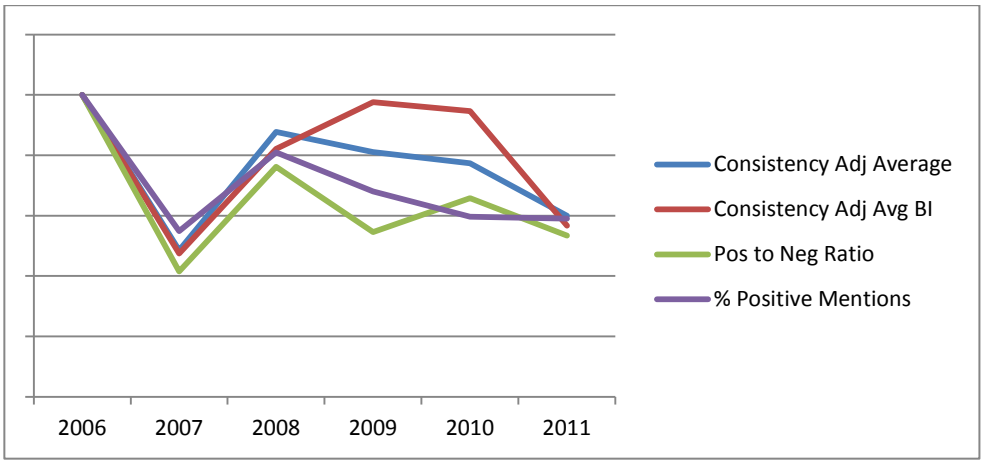

Figure 9. Hussey's Performance versus Sentiment

Figures 5, 6, and 9 show that Ponting, Sehwag, and Hussey have a decent correlation between their CAA and their percentage of positive sentiment mentions. The correlation between all four of Ponting's metrics is fairly strong.

Tendulkar's case is interesting (Figure 4). You see a huge spike in his CAA from 2009 to 2010. However, his sentiment showed a downward trend since 2009. The reason for this is that in 2010, Tendulkar played only two ODIs, and one of them happened to be the match in which he had his highest score (and one of the highest scores in the history of the game) of 200 not out. This score, combined with the fact that it was an inning in which he was not out, caused his CAA to soar in 2010; this is an outlier. If you look at his sentiment for 2010, you can see that it actually fell significantly from 2009. This trend is easily explained. In 2010, Tendulkar was rested from a number of ODI matches, and there was a lot of discussion about the fact that he was getting old and should retire. This potentially translated into an increase in negative sentiment. Despite this increase, he still had a high positive-to-negative sentiment ratio of close to 2.

Figure 7 shows that even though Dhoni's batting performance improved significantly between 2010 and 2011, his sentiment seems to have dipped during the same period. This could be because he is India's captain and India's performance as a team has fallen rapidly since the 2011 World Cup. As the captain, he could be taking heat for the team's poor performance.

CONCLUSION AND FUTURE WORK

Although in some cases a strong correlation is seen between the sentiment and the on-field batting performance of a player, it cannot be generalized. No single metric can comprehensively measure the performance. Multiple metrics should be considered together. Just looking at the overall sentiment toward a player might not be sufficient to truly

understand how the fans feel about him. More granular levels of sentiment need to be examined. Analyzing structured and unstructured data together provides greater value. This work can be extended along the following directions in the future:

- Using tools such as SAS Enterprise Miner to analyze a team's or a player's performance in order to understand the key metrics that influence their success.
- Collecting the unstructured data from an extended set of sources such as social media and online traditional media in England and Australia, where the game is very popular.
- Extending the unstructured data analysis by using a more granular sentiment analysis model in order to understand the sentiment around the different aspects of a player.

REFERENCES

Albright, Russ, and Lakkaraju, Praveen (2010). "Combining Knowledge and Data Mining to Understand Sentiment: A Practical Assessment of Approaches." *M2010 Data Mining Conference.* Available at: http://www.sas.com/reg/wp/corp/27999

Borooah, Vani K., and Mangan, John E. (2010). "The 'Bradman Class': An Exploration of Some Issues in the Evaluation of Batsmen for Test Matches, 1877–2006." *Journal of Quantitative Analysis in Sports*: Vol. 6, No. 3, Article 14. Available at**:** http://www.bepress.com/jqas/vol6/iss3/14. DOI**:** 10.2202/1559-0410.1201

Bull, Andy (2008). "Winning by Numbers." *Guardian* (London), May 8. Available at: http://www.guardian.co.uk/sport/2008/may/08/cricket1

ESPN. Cricinfo (2012). Available at: http://www.espncricinfo.com. Accessed on March 5, 2012.

Gini, C. (1912). "Variability e mutability; contributo allo studio delle distribuzioni e delle relazioni statistiche." Bologna: Tipografia di Paolo Cuppini.

Impact Index Cricket Pvt. Ltd. (2012). http://www.impactindexcricket.com/

Lange, Kathy, and Sethi, Saratendu. "What Are People Saying about Your Company, Your Products, or Your Brand?" *Proceedings of SAS Global Forum 2011*. Available at: http://support.sas.com/resources/papers/proceedings11/158-2011.pdf

ACKNOWLEDGMENTS

The authors are grateful to Anne Baxter and Ed Huddleston for their editorial contributions. The authors would also like to thank Murali Pagolu for his helpful discussions about the concepts presented in this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Praveen Lakkaraju
Enterprise: SAS Institute Inc.
Address: 10 Fawcett St.
City, State ZIP: Cambridge, MA 02138
Work Phone: 617-576-6800
E-mail: Praveen.Lakkaraju@sas.com

Name: Saratendu Sethi
Enterprise: SAS Institute Inc.
Address: 10 Fawcett St.
City, State ZIP: Cambridge, MA 02138
Work Phone: 617-576-6800
E-mail: Saratendu.Sethi@sas.com