Title	TweetCric: A Twitter-based Accountability Mechanism for Cricket
Author(s)	Younus, Arjumand; Qureshi, M. Atif; Aljohani, Naif R.; Greene, Derek; O'Mahony, Michael P.
Publication date	2017-06-01
Publication information	Web Engineering: 17th International Conference, ICWE 2017, Rome, Italy, June 5-8, 2017, Proceedings
Conference details	17th International Conference on Web Engineering (ICWE 2017), Rome, Italy, 5-8 June 2017
Series	Lecture Notes in Computer Science book series (volume 10360)
Publisher	Sage
Item record/more information	http://hdl.handle.net/10197/9168
Publisher's statement	The final publication is available at www.springerlink.com.
Publisher's version (DOI)	http://dx.doi.org/10.1007/978-3-319-60131-1_43

TweetCric: A Twitter-based Accountability Mechanism for Cricket

Arjumand Younus[1], M. Atif Qureshi[1], Naif R. Aljohani[2], Derek Greene[1], and Michael P. O'Mahony[1]

[1] Insight Centre for Data Analytics,
University College Dublin, Dublin, Ireland
[2] Faculty of Computing and Information Technology,
King Abdul Aziz University, Jeddah, Saudi Arabia
{firstname.lastname@ucd.ie and nraljohani@kau.edu.sa}

Abstract. This paper demonstrates a Web service called *TweetCric* to uncover cricket insights from Twitter with the aim of facilitating sports analysts and journalists. It essentially arranges crowdsourced Twitter data about a team in comprehensive visualizations by incorporating domain-specific approaches to sentiment analysis.

1 Introduction

Cricket is an international sport with a massive number of fans from around the world. Regions within South Asia in particular have a massive fan base, making the game analogous to a religion [1]. Over the years there have been cases of corruption within the game [3], and in many instances selection of cricketers for future matches is driven by politics [6]. Towards the aim of facilitating the "human in the loop" within cricket, this demonstration paper proposes a Twitter data aggregator designed to help sports analysts and journalists in deriving deeper insights into a game which in turn can lead towards better decision-making processes within cricket. Similar to some of the techniques deployed by the World Bank [2], *TweetCric* attempts to utilise big social media data for innovative analytics in cricket such as highlighting certain cricketers associated with negative sentiment on a frequent basis.

TweetCric essentially mines the crowdsourced Twitter data posted by cricket fans and presents it in an exploratory search interface. The interface provides interactive visualizations which supports the user by (1) extracting overall tweet volume activity for a team, (2) highlighting cricketers who receive more attention at different time intervals and (3) displaying sentiment expressed for a cricketer at different time intervals of the match.

In line with *TwitInfo* [5], our system allows a user to monitor activity peaks around an event (which in this case is a cricket game) along with associated sentiment information. However, *TweetCric* presents a complete snapshot of the event while also providing entity-specific sentiments to evaluate performance of various entities within the event[3].

[3] In the context of *TweetCric* an entity represents a cricketer

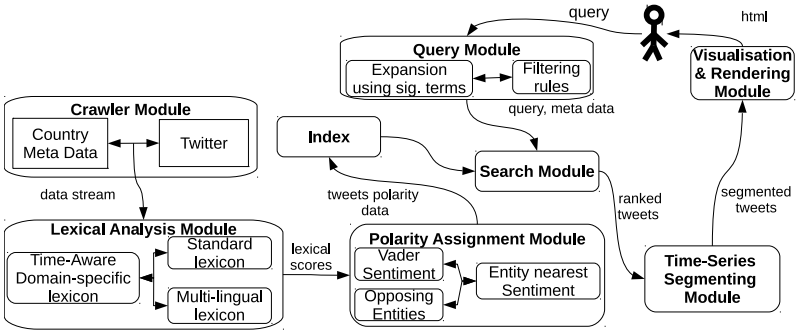

Fig. 1: Architecture of the *TweetCric* system.

2 System Architecture

Figure 1 shows an overview of the *TweetCric* architecture. The *crawler module* is responsible for back-end data acquisition, continuously collecting tweets from the live stream using the Twitter Streaming API[4]. The *crawler module* also gathers metadata (e.g., player names and their Twitter accounts, teams' management, match venues etc.) relating to countries that play cricket.

When the user issues a query, it is processed by the *query module* to produce a ranked list of relevant tweets. Note that within *TweetCric*, a query represents a certain player of interest (which we refer to as an entity) in the game for which the user of our system wishes to conduct a performance analysis. The query is automatically expanded through the use of significant terms appearing in the tweets, where significance is calculated by the chi-square test of independence. The chi-square test is able to detect term dependencies, and is hence useful in identifying useful terms to include together with an entity-based query. For example, the query "Shahid Afridi" is expanded with the term "Lala" which is a popular nickname of cricketer "Shahid Afridi". The query module also applies a set of filtering rules to eliminate noisy tweets[5]. The *time-series segmenting module* partitions the ranked list of tweets into segments which represent equal-sized time intervals between the timestamps of the first and last tweets among those retrieved for the query. Note that the distribution into equal-sized time intervals is a design choice motivated by the nature of cricket wherein a single time interval influences the entire match, and activity peaks are not as pronounced as in other sports. Further, this module also determines informative terms for each segment by calculating their importance using standard tf-idf weights. In the *polarity assignment module*, each tweet is assigned a polarity (sentiment) score as explained in the next section. Finally, the *visualisation and rendering module* generates the final HTML output in the form of visualizations and a ranked list of retrieved tweets.

[4] `https://dev.twitter.com`

[5] Tweets directed to political accounts, off-topic tweets such as related to showbiz, and tweets from known bots.

3 Domain-Specific Sentiment Analysis

TweetCric incorporates a domain-specific approach to sentiment analysis whereby the *lexical analysis module* and *polarity assignment module* (see Figure 1) work in combination with each other to incorporate algorithmic changes during the calculation of sentiment scores. Firstly, multi-lingual terms are added to a standard lexicon (the VADER lexicon [4]) to produce an extended lexicon. This is done after a manual inspection of partially non-English tweets (i.e., tweets posted in Indic languages but written in the Roman script) to introduce non-standard sentiment terms (65 in total which comprise 28 positive terms and 37 negative terms). Additionally, the *lexical analysis module* combines some cricket terminology with context-specific information to compute a final sentiment score. In this particular case, time is the significant contextual feature and we compute a sentiment score for cricket terms while taking into account timestamps. For example, terms such as "boundary" and "six", which denote the scoring of runs, are considered positive for the *batting* team but negative for the *fielding* team; similarly, terms such as "wicket" and "dismissal" are considered positive for the *fielding* team but negative for the *batting* team. Hence, the sentiment score of terms in the context-aware, domain-specific lexicon are sensitive to time implying positive scores for terms denoting *good batting* when our team of interest is the *batting* side, and negative scores for terms denoting *good batting* when our team of interest is the *fielding* side.

TweetCric performs sentiment aggregation at the following granularity levels: 1) player, 2) team, 3) game, and 4) time segment within a game. Once the set of tweets related to each of the above levels are identified, aggregation of individual tweet-level sentiments is then performed to produce an overall sentiment score.

In the case of entities (players), the following refinements are also considered. The *entity-nearest sentiment* scores a particular entity through a decay factor based on word distance (d) from the entity. Specifically, the scoring function is $Entity_{sentiment} = \sum_{w \in ExtLex} \frac{1}{d} w$ where d (set to 5 by experiment) measures the word distance between our entity of interest and a sentiment word w.

In addition, certain tweets mention two entities; as an example consider the tweet "Brilliant innings @imVkohli on a tough surface... Well done!!! Expected more from @SAfridiOfficial with the ball". Here, assume "SAfridiOfficial" is our entity of interest and "imVkohli" is a player from the opposing team. In these cases, we utilise information about such *opposing entities* (from within teams that are opposed to our team of interest[6]) to produce a sentiment score for our entity of interest. Here, the sentiment score of an opposing entity is simply subtracted to produce a sentiment score for our entity of interest.

Finally, we note that the opposing entities approach to sentiment can be further refined and that the approach has wider application – e.g. in the politics and technology domains, where opposing politicians and competing electronic products can be identified. These matters are left to in future work.

[6] The metadata about entities pertaining to different countries is obtained by the *crawling module* (see Figure 1).

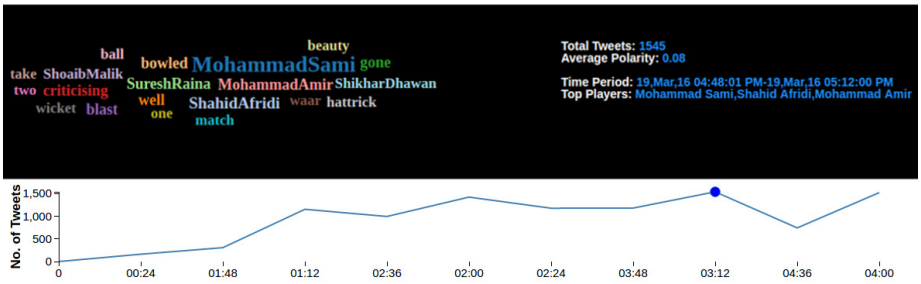

Fig. 2: *TweetCric* interface showing results for team "Pakistan" in the "India vs. Pakistan" World T20 game

4 Demonstration Plan

TweetCric encompasses a query-based exploratory UI; a full walkthrough of the system can be accessed at `http://209.159.151.130/tweetcric.html`. As mentioned in Section 2, the interface supports various modalities showing summary information for an entire game and a certain cricketer together with retrieved tweets. Furthermore, users are provided with navigation capabilities for significant entities (i.e., cricketers) at various points of the game. As an example, Figure 2 shows results for a crucial game between "India" and "Pakistan" with the query on cricketer "Muhammad Sami"; various points in the game can be clicked and each point highlights a word cloud summary.

References

1. Livin' on a prayer - espncricinfo . `http://www.espncricinfo.com/blogs/content/story/996281.html`.
2. N. A. Calderon, B. Fisher, J. Hemsley, B. Ceskavich, G. Jansen, R. Marciano, and V. L. Lemieux. Mixed-initiative social media analytics at the world bank: Observations of citizen sentiment in twitter data to explore "trust" of political actors and state institutions and its relationship to social protest. In *Big Data 2015*.
3. C. Davies. Match and spot-fixing: the challenges for the international cricket council. *Sports Law ejournal*, 2015:1–9, 2015.
4. C. J. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*, 2014.
5. A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *SIGCHI 2011*, 2011.
6. B. Radford and D. Hair. *Caught Out-Shocking Revelations of Corruption in International Cricket*. John Blake Publishing, 2012.