Sports Match Analysis using Social Media Mining

Domain : Data Mining

Alfin William
LTVE17MCA058
Roll No : 55

MCA S6

Overview

- Introduction
- Literature Review
- Problem Definition
- Motivation
- Objectives
- Proposed System
- Experimental Evaluation
- Data Sets
- Results & Discussion
- Future Scope
- Publications
- Conclusion

Introduction

- A live sports match conveys ***tremendous amount of information real-time*** in a fast pace.

- This information about the ***gameplay*** , ***individual player performance*** & the one of many *unique strategies* that led to success should be analysed and recorded.

- Conventional data analysing methods are ***costly and resource consuming*** , whereas combining the strength of Web data mining and Sports can yield better results and is economical in general.

Literature Review

AUTHOR	PAPER TITLE	YEAR OF PUBLISHING	METHODOLOGY
C. Xu, J. Wang, K. Wan, Y. Li, and L. Duan	Live sports event detection on broadcast video and web-casting text	2006	Using 4 modules: live text/video capturing, live text analysis, live video analysis, and live text/video alignment.
Young-seok kim and Mijung kim	'A Wisdom of Crowds': Social Media Mining for Soccer Match Analysis	2019	Web scraping & API of collection, Clustering algorithms for organising and retrieval of data
Nichols, J. Mahmud, and C. Drews	Summarizing sporting events using twitter	2012	Unsupervised algorithm for generating a textual summary of events & comparison with human results.

Problem Definition

Conventional Method

Current Sports team performance analysis include following methods:

- **Video data** :

1) **Video data** from on-field cameras divided into thousand of individual events and analysing player speed , performance , posession time.

- **Wearable devices :**

1) **Wearable devices** using technologies like GPS to capture performance based on running speed , cadence , distance , time etc

Drawbacks

Conventional methods drawbacks include:

- **Video data** processing & analysis are complicated , computationally burdensome and not exactly cost efficient.
- **Wearable devices** needs favourable conditions , needs constant monitoring , maintenance and risk of failure.

Motivation

- Largely untouched area of Web Data Mining.
- Scope of this project can be extended to other similar scenarios and applications that work with realtime human interaction data.
- To analyse the effectiveness of Complex Data Mining Algorithms in computing & extracting useful information from largely unstructured human interaction data.

Objectives

- Collect & analyse social media fan comments & ***extract logically top comments*** for performance improvement in future matches.
- Divide fan comments into ***mutiple clusters*** of unique topics to analyse the specific area of gameplay they mostly focused & evaluated about.
- Perform ***Sentiment analysis*** on fan comment data to evaluate fan satisfaction and overall emotional response to the match.

Proposed System

The Fan’s Perspective :

- Fans comes in all shapes and sizes : from *Daily workers* to experienced *Data scientists*.

- They closely follow their team and anticipate different use cases, deriving conclusions from it and sharing it real-time on social media.

- This is a largely ***untouched pool of data*** that can effectively convey useful information to team’s management.

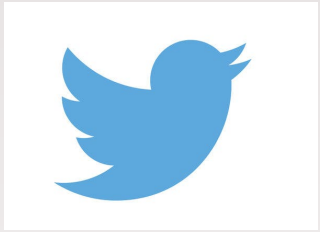
- Using this derived data , ***Future matches and Stratergies of teams*** can be improved and managed to further make efficient use of team’s resources.

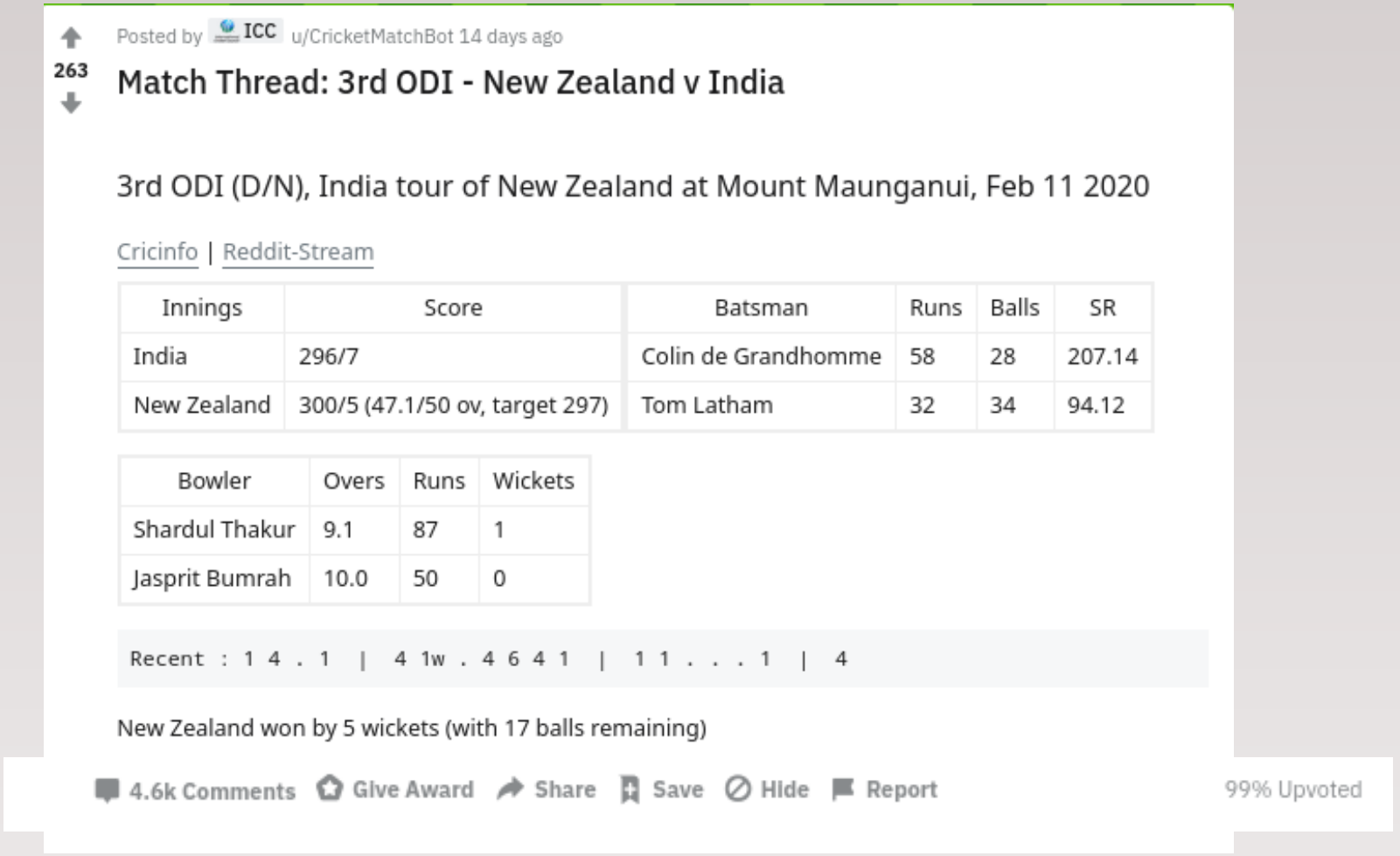
- Morever , this data collection & manipulation involves ***less economical and computational overhead*** compared to conventional methods.

System Design

I. <u>Data Collection Stage</u>

1) Web Data Scraping : Collecting user comments from social media sites using API's like **Twitter & Reddit API**. This requires Unique autentication keys and API Keys from client application.





263

Posted by ICC u/CricketMatchBot 14 days ago

Match Thread: 3rd ODI - New Zealand v India

3rd ODI (D/N), India tour of New Zealand at Mount Maunganui, Feb 11 2020

Cricinfo | Reddit-Stream

Innings	Score	Batsman	Runs	Balls	SR
India	296/7	Colin de Grandhomme	58	28	207.14
New Zealand	300/5 (47.1/50 ov, target 297)	Tom Latham	32	34	94.12

Bowler	Overs	Runs	Wickets
Shardul Thakur	9.1	87	1
Jasprit Bumrah	10.0	50	0

Recent : 1 4 . 1 | 4 1w . 4 6 4 1 | 1 1 . . . 1 | 4

New Zealand won by 5 wickets (with 17 balls remaining)

4.6k Comments Give Award Share Save Hide Report 99% Upvoted

- Sample Reddit Thread page for a Cricket match – This individual thread contained 4.6k Unique user comments.

2) Preprocessing Data : After collecting data , the relevant keywords are extracted using Data mining algorithm : -

- **TF – IDF Algorithm**
  [***Term frequency – Inverse Document frequency***]

a) Term frequency of individual relevant keywords in a document.

b) Compares the frequency with the frequency of that word in total collection of documents.

c) Predicts the relevance of that specific word using both frequencies.

II. Data Processing Stage

- **K – Means Clustering**

1) Derives ***homogeneous subgroups*** within the data corpus according to a similarity measure such as euclidean-based distance or correlation-based distance.

2)The results of clustering analysis help ODI experts identify ***similar fan requirements*** and in turn, specific fan requirements can be identified.

II. Data Processing Stage

- **Sentiment Analysis**

- *Natural Language Toolkit (NLTK)* of python was used for a ***Lexicon based method*** to sentiment analyse fan comments.

- Comments are classified into positive , negative or neutral by comparing it with labelled dataset of words.

III. Result Representation Stage

- **Ipywidgets**

1) This lightweight python GUI library can be used to accommodate the entire results of most relevant fan comments and their sentiment analysis and cluster classifications without causing computational and resource overhead.

Experimental Evaluation

1) Data Collection Phase

- Data set analysis revealed Reddit fan comments to be more ***critic*** and ***constructive*** than emotional reactions of twitter fans.

- The ***280 character limit*** of twitter proved to limit fans capability to express views.

- Since reddit followed ***thread formatting*** with comments and subcomments style , it provided content rich debate data.

2) Data Processing Phase

- *TF – IDF values* indicated fans mostly taking about a specific sports person or about a specific key event of the match.

- Higher weight scores contained country , sports person names & historical score values.

- K – Means Clustering proved strenuous since all data were from a similar domain.

- Distinct Clusters contained similar comments since fans talked about multiple issues in a single comment.

Screenshots

1) Reddit API results

	comment
0	So this is the worst series of Kohli like in y...
1	Saini ODI average at 53 😂
2	SANTNER OVERRATED\n\nNEESHAM OUTDATED\n\nLONG ...
3	Kohli running with Shaw is going to be interes...
4	[deleted]
...	...
4640	OK, thanks for your feedback.
4641	Fair enough. I just still try to remember they...
4642	Each to his own man
4643	😂😂
4644	Cool

[4645 rows x 1 columns]

2) TF – IDF words with weight scores

```
bhai_saheb  -  0.04099023818138398
15bfeqj1kl  -  0.02049511909069199
deba  -  0.02049511909069199
debates  -  0.02049511909069199
opened  -  0.04099023818138398
imvkohli  -  0.06148535727207597
sledge  -  0.04099023818138398
aussies  -  0.04099023818138398
matthewwade  -  0.04099023818138398
wicketkeeper  -  0.029164892066626372
australian  -  0.04099023818138398
words  -  0.04099023818138398
recent  -  0.04099023818138398
sportsflashes  -  0.02049511909069199
fn0uegxgss  -  0.02049511909069199
finale  -  0.02049511909069199
cricket  -  0.07291223016656594
h33qfkeafq  -  0.02049511909069199
```

2) Top fan Comments

Fans mostly talked about:

series	india	kohli	saini	like	bowling	rahul	cricket	bumrah

```
Top comments about : bowling

 Hit a century belting the piss out of a near full strength English bowling unit in a warm up.
 All 3 of them capable of building innings as well as hitting the big shots when needed"
 Things going so good for KL Rahul, he should open the bowling and see if he can get a wicket or two.
 "Only bowling is keeping grandmom in the side.
 Were finally not opening the bowling with Bennett, good choice
 I see it's India's time to attempt suicidal runs
 "So glad to see no 4&5 batsmen actually scoring runs
```

2) Cluster based top comments

Top Batting related Comments:

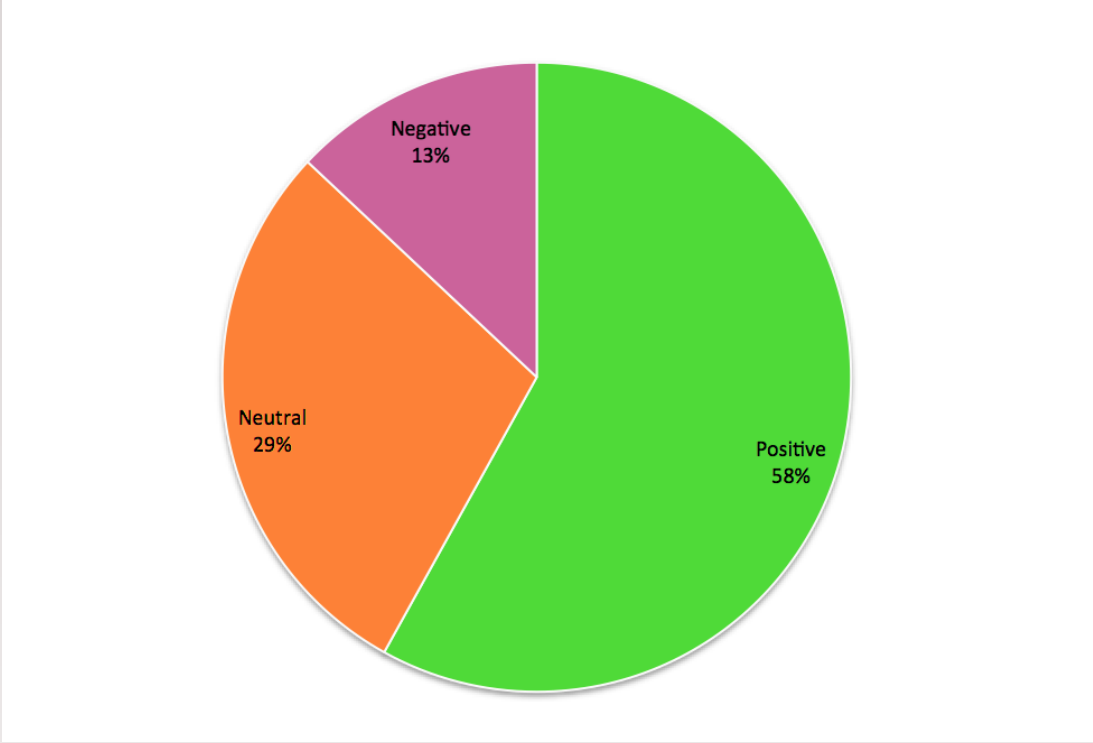
series	india	kohli	saini	like	bowling	rahul	cricket	bumrah

Top Bowling related Comments:

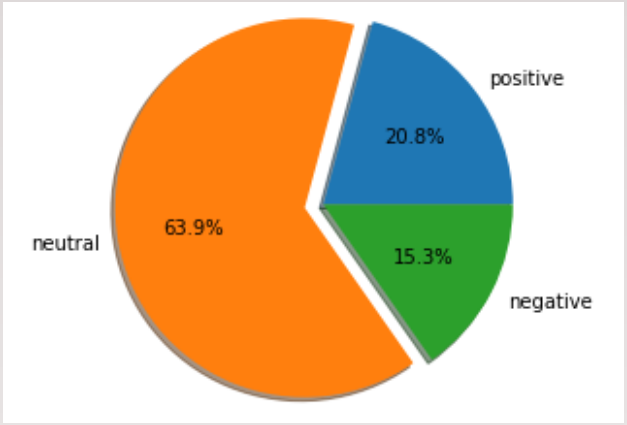
series	india	kohli	bowling	cricket	bumrah	thakur	wickets	team

2) **Sentiment Analysis**

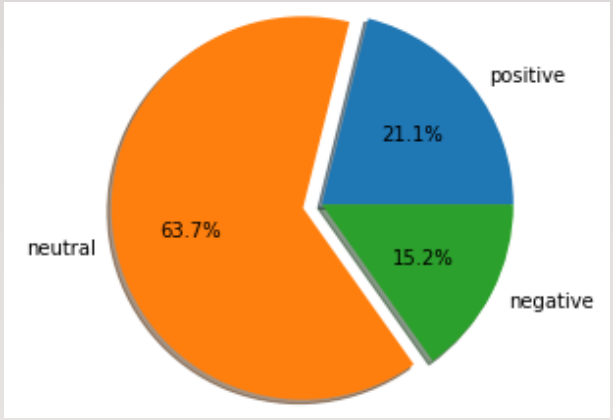
- Total fan comments



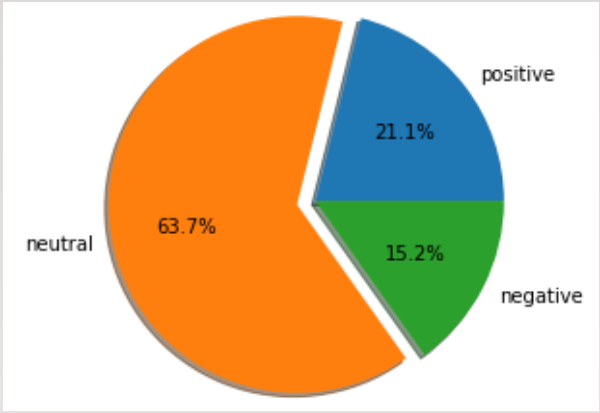
- Cluster Based Sentiment Analysis



Batting Cluster



Bowling Cluster



Fielding Cluster

Results & Discussion

- Data Scraping from social media sites were ***successful*** with use of REST API calls to clients.

- Content ranging around 5000 comments retrieved from reddit possessed more ***quality factor*** than twitter data.

- Preprocessing of text was imperative to weed out ***colloquial*** words , ***emojis*** and other noise from information.

- TF – IDF algorithm resulted ***accurate weight predictions*** with weights within the range of 0.23 and 0.01.

- K – Means Clustering with K value = 3 , seemed ***strenous*** due to similar domain of data.

- Sentiment Analysis was ***successfully completed*** with fan emotions captured and recorded for future analysis.

- The sentiments : Postive was above 50% in all cases followed by neutral emotion of around 15 – 30%.

Future Scope

- Expanding to other social medias using their official API’s.
- Automation and streamlining of the entire process.
- Realtime collection & computation using *Cloud computing*.
- Collection & storage of large data corpuses both historical data and present using *Big Data methologies*.

Conclusion

- The project was successfully completed with results recorded.
- Future enhancements were analysed and feasibility discussed.
- Use of social media data for sports match analysis was realised and proven using valid results.
- All processed data collected and stored in appropriate formats for future uses.