Sports Match Analysis using Social Media Mining

Domain : Data Mining

Alfin William
LTVE17MCA058
Roll No : 55

MCA S6

Overview

- Introduction
- Motivation
- Problem statement
- Objective
- Literature Review
- Design
- Methodology
- Progress of Work
- Timeline
- Future Scope
- Reference

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.

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.

Problem Statement

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.

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.

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.

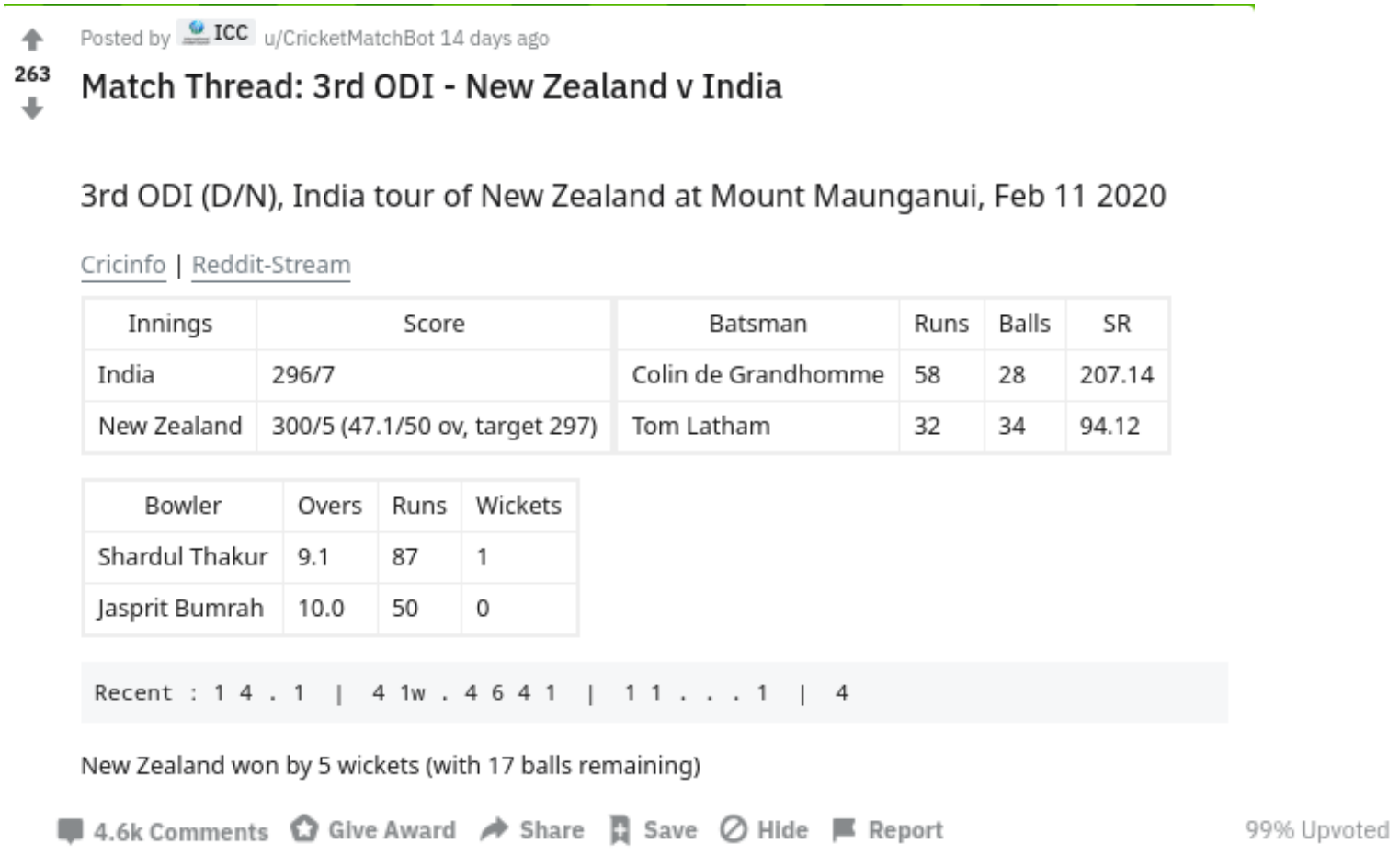
Literature Review

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.

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**

1) ***Multi-Nomial Naive Bayes Classification*** using scikit-learn is the optimal method to classify fan comments in this specific scenario.

III. Result Representation Stage

- **PyQt Framwork**

1) This lightweight python GUI framework 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.

Methodology

- Jupyter Notebook
- Anaconda Python Distribution
- K Means Clustering
- Multi-Nomial Naive Bayes Classification
- PyQt Framework

Progress Of Work

- Understanding the libraries and required dependencies.
- Setting up development environment.
- Collection of required dataset using web scraping.
- Preprocessing of collected data using stopwords and TF - IDF.

Timeline

- Complete Preprocessing of data [March 01 – 07]
- Starting Clustering of Processed data [March 09 – 17]
- Executing Data Mining Algorithms [March 17 – 25]
- Sentiment Analysis on Collected data [March 25 - 31]
- Creating a GUI environment via PyQt [March 31 – April 7]

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*.

References

- C. Xu, J. Wang, K. Wan, Y. Li, and L. Duan, “Live sports event detection based on broadcast video and web-casting text” in Proc. 14th ACM Int. Conf. Multimedia, 2006, pp. 221–230

- J. Nichols, J. Mahmud, and C. Drews, “Summarizing sporting events using twitter,” in Proc. ACM Int. Conf. Intell. User Interfaces, 2012, pp.189–198

- “A Wisdom of Crowds’: Social Media Mining for Soccer Match Analysis” by Young-seok Kim 1 , (member, IEEE), and Mijung Kim (member, ieee)