Engineering and Technology Journal e-ISSN: 2456-3358
Volume 05 Issue 08 August-2020, Page No.-730-735
DOI:10.47191/etj/v5i8.02, I.F. – 6.127
© 2020, ETJ

Knowledge Collection from Cricket Articles in English Newspapers

S. Srihari[1], N. Ilakiyaselvan[2]

[1]SCOPE Vellore Institute of Technology, Chennai Chennai, Tamilnadu, India
[2]Assistant Professor, SCOPE Vellore Institute of Technology Chennai India Chennai, Tamilnadu, India.

Abstract: Sentimental analysis is a hugely popular application of natural language processing. It has various uses and is mostly used to classify opinions. In this new age world AI and NLP has forced its way into every industry. The aim of our project is to use sentimental analysis in the domain of sports especially cricket. Based on a repository of positive and negative statements from newspaper articles we try to use them to perform analysis of players especially performance and popularity analysis. We also aim to create player profile pages which helps in structured and deeper analysis of cricket player.

I. INTRODUCTION

With time, our lives have become much faster paced than it used to be. Technology has a very big role in making it so. With everything coming to online platform nowadays, people are able to make the most of their time. There are a lot of things people have switched from offline to online mode. News has been no different. Digital viewership of news through the medium of Mobile apps, websites etc. has also increased a lot. Important knowledge about various players and various series could be gained by knowledge extraction from Sports articles in English newspapers.

Natural Language Processing, abbreviated as NLP, is a branch of artificial intelligence which deals with the interaction between computers and humans using the natural language [the way how humans communicate]. The ultimate objective of NLP is to read, decipher, understand, and make sense of the human languages in a manner that is valuable. Natural Language Processing is among the widely popular branches of Artificial Intelligence which has found its way into daily human life and has transformed our lives silently. One of the more popular applications of natural language processing include sentimental analysis. The simple definition of sentimental analysis is that sentimental analysis is the interpretation and classification of emotions within text data using NLP techniques.

Sentimental Analysis is hugely popular and has variety of applications. Few such applications include customer review classification as negative and positive which helps companies understand the user experience and reactions towards the product. There are also many other major applications. Every industry uses sentimental analysis for better understanding and analysis these days. In this project, by choosing the domain of sports (in specific cricket) we aim to perform performance and popularity analysis of various players involved in the sport with sentimental analysis being one of the contributing factors for the analysis.

Newspaper articles are sources of great information. The articles may be a game report, an opinion piece, an interview or some other article. These articles reflect at various points positively or negatively about each player. For eg – if a player is told to be playing brilliantly by the article it sends a positive impact. While if the player is written about being playing poorly it gives a negative impact. So this can help in determining the performance of a player over a period of time. Talking about popularity, a player like Virat Kohli is always in the spotlight because he is being talked about in one article or the other almost every day. So based on how much a person/player is being talked about/written about one can make an analysis of how much a player is important and popular.

Sentimental analysis is the interpretation and classification of emotions with textual data by the usage of textual analysis techniques. It is extremely useful in various domains, especially businesses for identifying customer sentiment towards products, brands or services in online conversations and feedback. Even though sentimental analysis is a very useful and popular application the data obtained from sentimental analysis /could be used in many applications. Our aim to review few such applications where the data obtained from sentimental analysis is used. For example, it could be used for stock market predictions, predictions in sports and other such applications.

Sports based analysis are always useful in a lot of ways. Team managements can actively analyse a player through a combination of these techniques and can make out a public opinion / consensus on how a player is performing. Club and Team owners through these types of analysis can benefit from these kinds of analysis because it helps them to garner much needed public opinion and support required to run the respective club/team. Sponsors and Advertisers can use the analysis actively to mutually benefit from partnerships with players by actively using related analysis techniques. This can be also used in new age fantasy leagues wherein these kinds of analysis can help participants actively choose and represent their options wisely.

II. LITERATURE SURVEY

In Paper [1] by Srivastava R et al., Capital market forecasting is aimed to be done by using sentimental analysis. Whenever a customer invests into stock markets, he aims to attain higher profits in a short period of time but due to minimal knowledge the process becomes very difficult. It also becomes dangerous and could prove to be fatal for the customer. So for facilitating better understanding of the market the system of paper [1] uses sentimental analysis.

Stock Market Prediction involves the requirement of previous years of stock data of a particular company, which could be by retrieved from finance websites and other appropriate platforms. The data is then preprocessed and a dataframe is formed. The future price of stock is obtained by exponentially weighting the data (close price) which is done by first selecting the number of days of data to be taken as input then giving the highest weight to the latest data and decreasing the weight to older ones with farthest data getting the lowest weight. Future price of stock is obtained by simple averaging the data (closing price) is done by initially selecting the number of days of data to be taken as input and then adding all the data and then performing their average. For predicting next day’s polarity of a stock i.e whether the stock price of a company will go up or down classification techniques can be used. KNN / SVM was used for classification purposes. For predicting market sentiment of the firm, the authors of [1] have used sentiment analysis which uses natural language processing and text analysis to identify and extract required information from the data source. Hence, the company’s next day close price, its polarity and the market sentiment of the company were all predicted. All three of these information can be used by the investor for getting a better insight of the next day’s trends before buying or selling of company’s stocks.

In Paper [2] by Singh N et al., a system is proposed wherein production prediction is done based on news using sentimental analysis. This is done due to the fact that the market is changing rapidly. The industry needs to satisfy the needs of the users and at the same time make correct business decisions .News from the online sources is mined and is combined along with sentiment analysis for producing the best suitable suggestions. The recommendation which is then given to the production based companies helps them decide to speed up or speed down the production so that one could take the optimum benefits from the market. The association analysis is performed to show the words which are contextually associated with each other. The news article and its complete analysis is performed and verified. In the proposed work four different phases are recommended. Every phase is dependent on the output of the phase before and cannot be parallelized. The authors use the following features:

Relevance: **Every news is NOT equally relevant to everyone**. So there is a need to filter out irrelevant news. The relevance is calculated based on the keywords associated with the organization.

Strength Analysis: The detailed news is fetched and Nouns, Verbs and Adjectives are extracted respectively. The strength of news on particular business is made out by finding impact of one word on other.

- **Impact Analysis**: The impact on businesses by news can be bipolar. The system uses association analysis and sentimental analysis to calculate the overall impact.

The production speed is calculated with the help of Conf(PS) and Conf(NS) which is received as the result of Impact Analysis phase. Recommendations are generated by the decision support system proposed for various industries and businesses.

In Paper [3] by Kim J et al., prediction of stock prices are done based on sentimental analysis of news articles. Recently research on stock market integrated with AI has seen a surge. In this proposed work through sentimental analysis, one can obtain the positive index of news articles for each date. By analyzing the correlation b/w positive index value and the stock return value, one can confirm the utility and possibility of the sentimental analysis in stock market. The proposed system requires data like newspaper articles and stock prices for each date. For stock prices next day stock price (NSP) and stock price return (SPR) are required. When it comes to news articles, they are sorted chronologically and stop words are removed. Then they are fed to morphological analyzer that makes word noun. To build a sentiment dictionary, the authors made use of the frequency, positive and positive index (PI) of the words. [No of times occurred, no of times the word occurred on the day when stock price increases, positive value/frequency].

$$\text{word}(i,j) = \text{The number of times a word appeared in a news article}$$

$$\text{frequency}(i) = \sum_{j=1}^{n} word(i,j)$$

$$\text{NSP}(j) = \begin{cases} 1: \text{If the next day's stock price rises after the article j is posted} \\ 0: \text{next days' stock price decreases or same after the article is posted} \end{cases}$$

$$\text{Positive}(i) = \sum_{j=1}^{n} \{word(i,j) \times NSP(j)\}$$

$$\text{PI}(i) = \frac{\sum_{j=1}^{n}\{word(i,j) \times NSP(j)\}}{\sum_{j=1}^{n} word(i,j)}$$

In [4], a paper bhy Khatri SK and Srivastava sentimental analysis is used in prediction of stock market investment. In the proposed work, sentimental analysis is performed on the data retrieved from Twitter and Stock Twits. The data is analyzed for computing the mood of user's comment. These comments are categorized into four categories which are happy, up, down and rejected. The polarity index also with market data is supplied to an artificial neural network for predicting the results.

User tweets are oftentimes useful in knowing user opinion on brands. Tweets about various top companies were collected [Apple, Microsoft, Oracle, Google and Facebook].The tweets were classified based on polarity [happy, up, down, rejected]. The overall index values and the market data of that firm is fed as the inputs which are passed to an artificial neural network [ANN] to train and predict. This input data for an artificial neural network is in form of a csv workbook where first four columns determine the market data indexes and last column determine the overall index value of sentiments for that company. After this is done closing prices are compared to tell which company a person should invest.

In [5] a paper by Jagdish Chandra patni et al., the authors motivation is to generate a system that can harness the wisdom of crowds using the sentiment information from Twitter to make match predictions. The tweets were used as the input and through NLP techniques predictions of outcome of cricket matches was done.

Tweets of 3 different matches were captured in 3 different categories buckets The size of each category which is called as a bucket is around 8,000 to 10,000 tweets which are further sub-divided in data store having maximum of 200 tweets each just to be safe in case if any debugging is required on raw data in future. Each tweet has all attributes like tweet Id, time stamp, tweet text etc available in JSON format in the data set.

Preprocessing on the tweet buckets were done. A model was trained to make a sentiment analysis classification model [Sad, Neutral, Happy, and Ecstatic]. . The tagged named entities from the training file are fed as input to the classification module and on the top of it defined were Gazettes for all of the four named entities which consist of all the players, locations, teams and venue names along with the acronyms and pet names of the players. The classification module then gives the separate named entities for all 4 classes. The output is processed for forming a JSON file which is then given as input to Data Driven Document module for graphical representation.

The tweets were tagged with different events (TOSS, BOUNDARY, WICKET, RESULT, and MILESTONE).One tweet was randomly selected from each event and an initialization. K- Means clustering is done. Cluster was formed with centroids mapping to the vector representation of the tf-idf values of tokens filtered. Features were extracted to remove stopwords, and few tokens matching POS tags (!, ', #,@, P). This feature representation forms a bag of words and created was a tf-idf vector representation for each tweet and applied was the k-means algorithm with the above initialization setup. For similarity also calculated was the cosine distance between the tweets and matched tweets to nearest clusters. The convergence of the algorithm takes place at around at 6th iteration for around 8000 tweets and k=5 (number of unique events). Off-the shelf techniques has been used for exploring random clusters. There were few interesting explored which formed clusters on some named entities. This approach is based on finding the peaks in the frequency of tweets and applying the technique of summarization only on the tweets in the window of peaks. Finally merging all the intermediate summaries to form a match summary.

The authors formed eight models which were Subjective negative, Objective negative, Subjective positive, Objective positive, All subjective, All Objective, All negative, All Positive objective. The authors concluded that Sentiment is a predictor of match and tournament outcomes and all models beat random chance. Their analysis also found that Subjective positive performed better but its chances were hurt by draw outcomes

III. PROPOSED SYSTEM

We undertook the task of creating a system that "takes an input of corpus of sports articles of a tournament and do the performance analysis and popularity analysis" automatically.

An end user might desire an automated overview like presentation of the main points made in a single review or how opinion changes time to time over a period during the tournaments. To meet the requirement used is the 5W task which seeks to extract the semantic constituents in a natural language sentence: Who, What, When, Where and Why. The visualization system facilitates users for generating sentiment tracking with textual summary and sentiment polarity wise graph based on any dimension or combination of dimensions as they want i.e."Who" are the actors and "What" are their sentiment regarding some topic, changes in sentiment during -"When" and "Where" and the reasons for change in sentiment - "Why".

The project focuses on applying NLP techniques for doing analysis of relevant, unstructured data of tweets on a given cricket match day. The solution can be used by marketing companies for incrementing user engagement on the targeted website. Also, the researchers of Human Behaviour can make a difference by analyzing the pattern changes in the human reactions during the various events of the game. The business analysts can monetize the trends for increasing the website traffic. Wrapper project of this Streaming API was used to build a real time system.

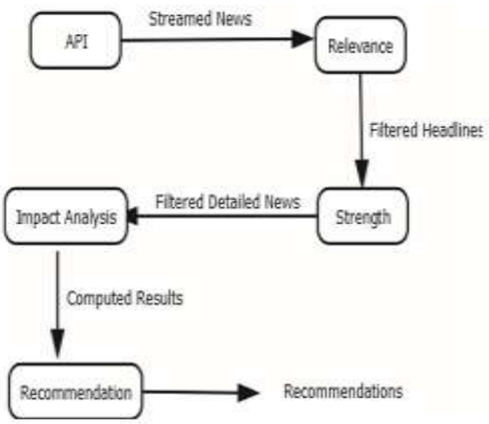


Fig. 1. Framework of Proposed Work

IV. METHODOLOGY

[a] Dataset creation and Pre-processing

For this purpose we took a set of 284 sentences and also took the polarity of the sentences [positive and negative]. We take the words in these sentences and tokenize them by word.

We identified a set of POS[Parts of Speech] tags which contribute to the statements being positive and negative. These tags are listed below.

JJ' - adjective
'JJR' – adjective comparitive
'JJS' – adjective superlative
'NN' - Noun
'NNS' – Noun Plural
'RB' – Adverb
'RBR' – Adverb comparative
'RBS' – Adverb superlative
'VB' –Verb
'VBD' – Verb Past tense
'VBG' – Verb Gerund
'VBN' – Verb past participle
'VBZ' – Verb 3rd person

We performed PoS tagging of the sentences and if any one of these tags occurred in the sentence, and if any one tag occurs, we stem/lemmatize the word and make the word part of a new sentence which consists of processed words in the same order as the original sentence**.**

[b] Machine learning model training for Sentimental analysis

From the newly framed sentences and already known polarity of these sentences we are going to train the model for sentimental analysis. Using the help of TF-IDF Vectorizer we vectorize it to transform it into the vector space model. Then we train the model with the help the vector space values as the input and the polarity as the required output. We use k nearest neighbours algorithm where the model uses 4 nearest neighbours i.e k = 4 to determine the polarity of the presented sentence.

[c] Web Scraping and parsing

We took a set of urls, all of those pertaining to the ICC 2019 World Cup from the Deccan Chronicle website. We scraped each of the url and then parse it to get the actual content of the web page.

[d] Co-reference resolution and Pronoun substitution

We took the content of the web page and scraped it and parsed to fin the content of the webpage. Then we did the process of coreference resolution. Consider the paragraph as follows.

"Ben Stokes produces a stunner at Headingley. He produced a lifetime innings to keep the England side alive at the Ashes Series." We can easily recognize the first sentence to be related to Ben Stokes based on similarity. But in the second sentence we need to recognize that he refers to Ben Stokes. So we use the process of coreference resolution to find that he refers to Ben Stokes and we substitute the 'he' in the second sentence by Ben Stokes. This step is necessary to find the sentences related to the player.

[e] Tokenize the content into sentences and find player (query) related sentences

After pronoun substitution we tokenize the substituted content into sentences by sentence tokenizer. Then we take the player name as the query and using that we find the related sentences based on the similarity between the query and the sentences.

[f] Performance analysis [Cumulative]

We take the urls for each day and parse the content after scraping the urls. After doing the process of coreference resolution and pronoun substitution and finding the player related sentences for articles of each day we classify it with the pre trained algorithm for classifying it to be positive or negative. After this process we find a cumulative score for the player by assigning +1 for a positive sentence and -1 for a negative sentence. After taking into all player related sentences in a day we give a cumulative score which we add it to the previous day score to find the overall performance value of the player till that particular day [from the first day of consideration].We present the popularity in a neatly presented scatter chart.

[g] Popularity analysis

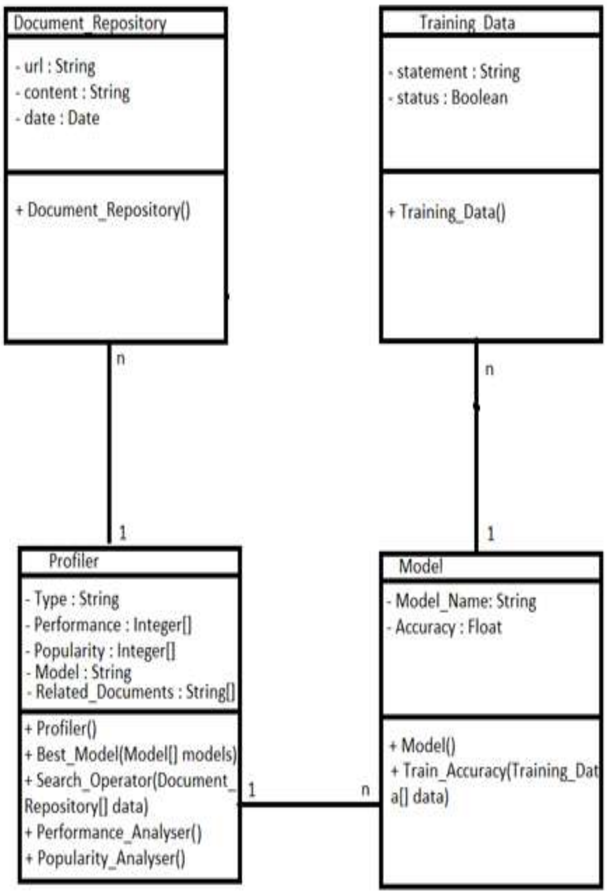
For popularity analysis we take the number of related urls to a particular player. More the number of articles more the amount of popularity/buzz is accounted for. The popularity is presented in a neat scatter chart.

[h] Day by Day Performance Analysis [Presented as Opinion Analysis for distinction]

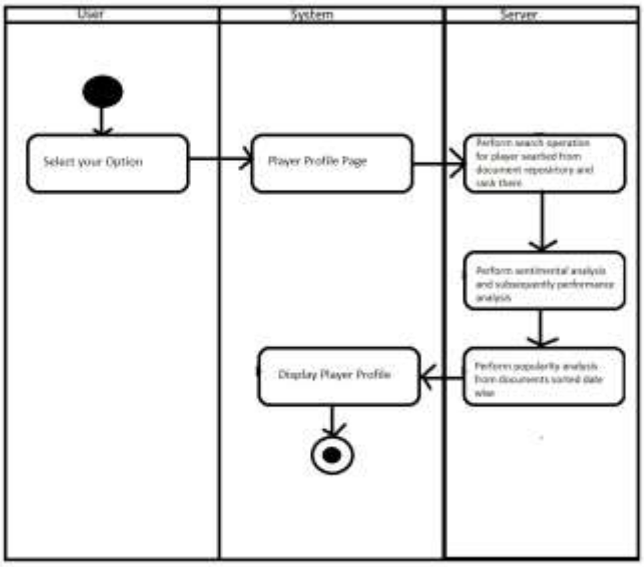
We take the no of articles presented on the player on a particular day and popularity score for that particular day. We divide the latter by the former to get the performance value/opinion value of/on that player that partilcar day. We have classified this into 3 segments : <0 a negative, between 0 and 1 as mixed and >=1 as positive performance/opinion.

V. DESIGN DIAGRAMS

The following is the Class diagram of ML training section of our system..



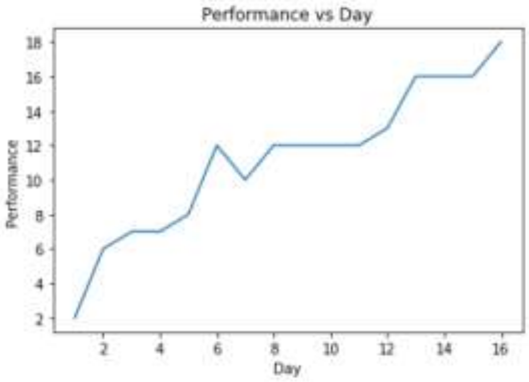
The following is the Sequence diagram of our system


User
System
Server
Select your Option
Player Profile Page
Perform search operation for player searhed from document repository and rank them
Perform sentimental analysis and subsequently performance analysis
Perform popularity analysis from documents sorted date wise
Display Player Profile

VI. RESULTS AND SCREENSHOTS

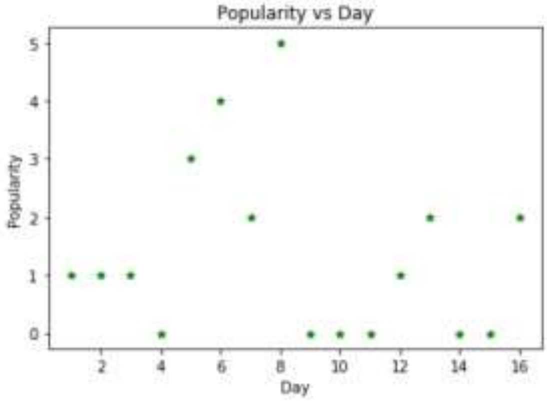
Here we are displaying the performance analysis of famous South African Cricketer Mr. Faf Du plessis. X-axis represents the days count and Y-axis represents the score of the player.

Diag1. FAF Du Plessis performance analysis



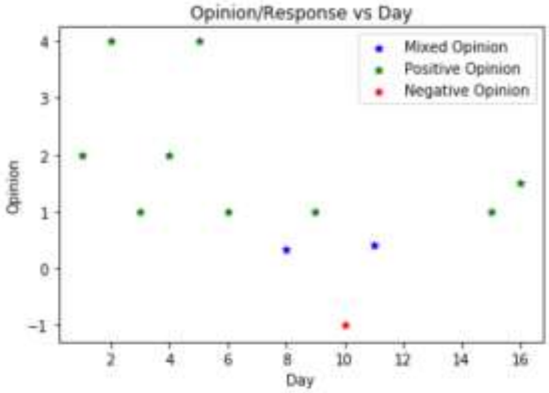
Diag2. FAF Du Plessis popularity

Here the popularity score of FAF Du Plessis is done. His popularity reached a peak on Day 8 nad on days 9,10,11,14,15 it was at its lowest**.**



Diag 3. Opinion analysis on FAF Du Plessis performance

The results of opinions about FAF Du Plessis throughout the tournament.



VII. CONCLUSION

Thus a successful system was developed for performing popularity and performance analysis of players and a player based search engine and thus a player profiler web application. This involved various NLP techniques and will be useful for various sports related business activities [for example deciding the value of a player for selection for IPL purposes].

VIII. REFERENCES

1. Srivastava R, Agarwal S, Garg D, Patni JC. Capital market forecasting by using sentimental analysis. In2016 2nd International Conference on Next Generation Computing Technologies (NGCT) 2016 Oct 14 (pp. 09-12). IEEE
2. Singh N, Tripathi A, Kumar V. Production Prediction based on News using Sentimental Analysis. In2019 4th International Conference on Information Systems and Computer Networks (ISCON) 2019 Nov 21 (pp. 32-36). IEEE.
3. Kim J, Seo J, Lee M, Seok J. Stock Price Prediction Through the Sentimental Analysis of News Articles. In2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN) 2019 Jul 2 (pp. 700-702). IEEE.
4. Khatri SK, Srivastava A. Using sentimental analysis in prediction of stock market investment. In2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO) 2016 Sep 7 (pp. 566-569). IEEE.
5. Jagdish Chandra Patni, Ravi Tomar, Mahendra Singh Aswal, Ankur Dumka. Prediction Based On Sentiment Analysis. International journal of Current Engineering and Scientific Research Vol 4, Issue 10, Oct 2017