PalArch's Journal of Archaeology of Egypt / Egyptology

ANALYSIS OF OPINIONATED TEXTS ON IPL 2020 MATCHES USING SOCIAL MEDIA DATA

Dr. Sakthi Kumaresh[1], Muskaan Shah[2], Samyuktha Sathyanarayan[3]

[1]Associate Professor, Department of Computer science, M.O.P. Vaishnav College for Women (Autonomous), Chennai, India

[2]UG Student, BCA,M.O.P. Vaishnav College for Women (Autonomous), Chennai, India.

[3]UG Student, BCA,M.O.P. Vaishnav College for Women (Autonomous), Chennai, India.

[1]sakthi.kma@gmail.com, [2]hmuskaan11@gmail.com ,[3]samyukthachess@gmail.com

Dr. Sakthi Kumaresh, Muskaan Shah, Samyuktha Sathyanarayan, ANALYSIS OF OPINIONATED TEXTS ON IPL 2020 MATCHES USING SOCIAL MEDIA DATA– Palarch's Journal of Archaeology of Egypt/Egyptology 17(7) (2020), ISSN 1567-214X.

Keywords:Opinion mining, Indian Premier League (IPL), Sentiments, Twitter, data analysis, Criciket

ABSTRACT

In India, The Indian Premier League (IPL) is a prominentand prestigious 20 overs cricket game conceptualizedwhere interstate teamswith a small representation of players from across countriescompete against each other to win the coveted trophy. Many leading icons from industry, films and sports enthusiasts own these teams with huge financial investment. Over the years IPL has become the largest revenue generating sports event in India. It has also brought about a sense of identity and belonging towards teamsas teams are named after cities where Indians reside. Usually,the matches are played in India at different locations. But, in 2020, because of Covid-19, the match venues were shifted to UAE and played without any audience. Given this context, public opinion about matches, cricketers and performance of teams are not only given out using live commentary but are usually expressed in the form of comments, tweets and opinion on several social media platforms. In this paper, Twitter posts (tweets) that relates to IPL teams and feedback of matches have been studiedand classifiedas positive and negative sentiments for understanding the sentiment and views of people about the different teams. Opinion mining has also been used to infer if people are happy watching the match at home or if they indeed want to go to the stadium to watch the match.

1. INTRODUCTION

It is said that three things unite India—Cricket, Bollywood and the English language. Indian Premiere League (IPL) is the heart of cricket fans with the 13th edition being conducted in 2020with eight teams representing eight different cities of India. The teams not only have Indian players but also have foreign players from various countries, except Pakistan.

This league acts as a platform for young cricketers to showcase their talent as well as a platform for entertainment. The format of IPL is such that every team plays two matches against every other and the top four teams in the points table qualify for the payoffs. The eight different teams who are currently playing in this season are Chennai Super Kings (CSK), Kolkata Knight Riders (KKR), Delhi Capitals (DC), Mumbai Indians (MI), Rajasthan

Royals (RR), Royal Challengers Bangalore (RCB) and Sunrisers Hyderabad (SRH), Kings XI Punjab (KXIP).

Usually, this league happens during the months of May and June in multiple cities with teams are competing against each other. But, because of the Covid-19 situation in 2020, IPL started late in September. The matches are held in UAE on three different grounds—The Dubai International Cricket Stadium, The Sheikh Zayed Cricket Stadium and the Sharjah Cricket Stadium. These matches are held with full safety and precautionary measures without any spectators.There was also this concept of bio-bubble where all the cricketers are kept in quarantine for the first 14 days after their arrival in UAE.

Millions of people express their opinions and views on a wide range of topics via microblogging websites and thousands tend to post their opinions and comments on Twitter and other social channels. To understand the sentiments and different points of view of the public, data which is in the form of posts has to be extracted from social media sites likeTwitter.

The analysis of the data based on the people's opinion is known as sentiment analysis. Text mining techniques and Natural Language processing (NLP) are used to categorize the emotions as positive, negative and neutral within text data. This process is known as polarity checking. Sentiment analysis enables us to interpret the text data based on the polarity checking. It lets businesses and individuals to identify the sentiments of their customers towards a particular product, brand or service when the feedback is taken online. It uses Natural language processing (NLP), biometrics and text analysis techniques to systematically extract and classify information.

The rise of social media such as blogs and social networking sites such as Twitter and Facebook have stimulated interest in sentiment analysis. With the increase in ratings and reviews, online opinion has become a benefit for businesses looking to market their products and identify new opportunities. In this paper, we use sentiment analysis in Twitter tweets to find out the sentiment of the people for a particular

team after the team has played the match. The analysis is done using Python to check how the fans react if a particular team loses or wins a match on a particular day.

Process followed by a basic sentiment analysis of text:

1. Breaking down the given text document under analysis into its constituent parts. The given text can be broken down into sentences, phrases and tokens.
2. Identifying every sentence that holds sentiments or a sentiment.
3. Assigning a score of sentiment otherwise called as a sentiment score to every phrase/sentence and component (-1 to +1)
4. Optional: Combining these scores of various phrases and components for multi-layered sentiment analysis

Many people consider opinion mining and sentiment analysis to be the same. But there are nuances and miniscule differences between the two. In the case of opinion mining, only public opinion and public mood is mined, but sentiment analysis is a deeper domain. Opinion mining is the study of people's emotions, opinions and appraisals towards entities, individuals, events, issues and their attributes. In sentiment analysis, emotions of the public are hidden in their messages and are assessed using sentiment analysis tools and algorithms. These emotions may be sad emotion, happiness, joy, anger, anxiety, etc.

2. LITERATURE REVIEW

Natural Language Processing (NLP) can be applied to do sentiment analysis. NLP task can be applied at many stages of coarseness. There are many ways of conducting sentiment analysis with well-known and traditional methods.

[3] Bingwei Liu et al have evaluated the scalability of Naive Bayes classifier (NBC) in large datasets in their paper. In place of a standard library,fine-grain control of the analysis procedure is executed by the NBC. The result tells us that the accuracy of NBC is improved as the dataset size increases and it is proved to be by 82%. This can also be used to scale up to do an analysis on the sentiment or feelings of millions of movie reviews with a constant increase in throughput.[4] Songbo Tan et al proposed Adapted Naïve Bayes (ANB) which uses the weighted transfer version of Naive Bayes. ANB is used to gain and receive knowledge from the new data domain[5] Lina L. Dhande1 et al combined the Naive Bayes and Neural Network classifier for sentiment analysis. The review is classified into positive or negative sentiments using classifiers. The accuracy of sentiment analysis can beincreased upto 80.65% by combining Neural networks with Naive Bayes classifier for unigram feature on movie review dataset. [6] BoyiXieet al featured analysis that revealed that the most crucialcharacteristics are those that shared or combined the previousdivergence of words of vocabulary and

their parts-of-speech tags by investigating two kinds of models: tree kernel and feature based models.[7] Andrew L. Maas et al presented a model that used a blend of supervised and unsupervised methods to learn rich sentiment content word vectors capturing semantic term–document information as well.In [7], the authors have proposed a model that leverages multi-dimensional or continuous sentimental information as well as non-sentimental comments. The model was useful in finding document level polarity annotations from many online documents. The model was evaluated using sentiment corpora and they came out with the results that their model out-performed other existing models for sentiment analysis.

[8] Alexander Pak et al took a basic approach to collect and classify tweets into three major categories like (i) tweets with emoticon queries such as ":-)", ":)", "=)" indicate positive emotion or happiness (ii) tweets with ": - (", ":(", "=(", ";(" indicate negative opinions or dislikes, and (iii) comments posted by newspaper accounts are considered objective or neutral. The authors also discussed how to retrieve data from a corpus for sentiment analysis and opinion mining automatically. Experimental results showed that their proposed techniques are better than the previously proposed methods.

To check whether an expression is neutral or polar, [9] Theresa Wilson et al presented a new approach to phrase-level sentiment analysis. Contextual polarity for a large subset of sentiment expressions can be identified using an automated approach that is provided in the study. Experimental evaluation of this approach helped the authors [9] to achieve better results than the baseline expectation.

To extract sentiments associated with polarities for specific subjects from a document, instead of classifying the whole document into positive and negative,[10] Tetsuya Nasukawa et al illustrated a sentiment analysis approach with a syntactic parser and sentiment lexicon.The prototype system achieved higher precision (around 75-95% depending on the data) in finding sentiments within Web pages and news articles. [1] I. Rishdiscussed about the Naïve Bayes classifiers in his paper. He showed that the Naive Bayes classifier is not directly correlated with the degree of feature dependencies measured as the class conditional mutual information between the features. The paper also demonstrates that naive Bayes works well for certain nearly functional feature dependencies, thus reaching its best performance in two opposite cases: completely independent features and functionally dependent features. [2]Mohammed J. Islam et al applied the Naive Bayes classifier to credit card approval testing data set and found that there is 12.43 (%) error of misclassification.

The authors presented vital options that accomplish a major gain over a unigram baseline. Additionally, they had a tendency to explore a distinct technique of data representation and report significant enhancements over the unigram models. The size of the data allowed them to carry out cross validation experiments and check for the variance in performance of the classifier across folds. Finally, they have given out their experimental outcome, which showed the precision in examining

the sentimental state of Facebook users, using the Naive Bayes Classifier, is very high [11].

The authors conveyed that using emoticons as noisy labels for training data is a powerful way to execute distant supervised learning. This paper also tells the pre-processing steps required in order to attain high correctness. They have shown the outcome of machine learning algorithms for classifying the sentiment of Twitter tweets using distant supervision [12]. This paper talks about the fourth year of the" Sentiment Analysis in Twitter Task". SemEval-2016 Task 4 contains five subtasks, three of which serve a remarkable departure from preceding editions. The three new subtasks focus on two variants of the primary "sentiment classification in Twitter" task. The first variant confers an ordinal character to the classification task. The second variant concentrated on the accurate approximation of the prevalence of each class of interest [13]

The authors' experiments show that part-of-speech properties may not be helpful for sentiment analysis in the microblogging domain. Using hashtags (#) to gather training data proved useful [14]. Their experiments showed that when microblogging features are incorporated, the benefit of emoticon training data is lessened.

In [15], they have focused to handle the difficulty of sentiment polarity categorization, which is one of the foundational issues of sentiment analysis. Test for both sentence-level categorization and review-level categorization were carried out with favourable outcome.

3. METHODOLOGY

Sentiment analysis is used to extract particular information in source material by applying numerous techniques such as computational linguistics, Natural Language Processing (NLP), and text analysis to classify the polarity of the opinion. This paper illustrates the research area of sentiment analysis on the tweets pertaining to IPL matches.Fig. 1 displays the process of sentiment analysis carried out

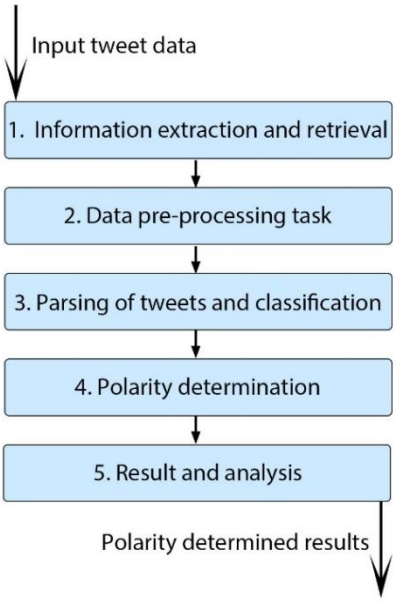

Figure 1- Steps involved in sentiment analysis.

3.1 Information extraction and retrieval

3.1.1. Installation and importing of the required packages

Tweedy: Thousands of tweets with particular hashtags are collected using the tweedy library of Python.Tweepy is open-sourced which is hosted on GitHub. It enables Python to communicate with Twitter platform and use its API.
TextBlob: Textblob is the python library for processing textual data.

3.1.2 Creating a constructor for the class which authenticates the Twitter client.

To do the sentiment analysis of Twitter tweets, a request has to be made for a Twitter API account. Twitter's API (Application programming interface) contains a set of Application's programming interface's or API's that can be grouped to create an application for a business. A customized application for business can be created using a set of Application's programming interface's (API's) using Twitter API. This is done so that it provides with an option of creating automatic tweets and one can benefit from the data offered by this social network. It is similar to a solution that helps a company respond to customer feedback on Twitter. Once a Twitter API is approved, few unique tokens are given to access the Twitter data. These unique tokens are authenticated in this step. If the API credentials are incorrect, an error is displayed.

3.2 Pre-process data

Data pre-processing is a data mining methodthatconverts raw data into ancomprehensible and logical structure. Real-world data is usually uneven, inadequate, lacking in certain behaviors, and tends to contain many errors.
In this step, we create a utility function to clean tweet text by removing links and special characters

Table 1- Python query showing the pre-processing of data

```
return ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-
z \t])|(\w+:\/\/\S+)", " ", tweet).split())
```

3.3 Parsing of tweets and classification

A function is created to fetch tweets and to do parsing of tweets. This step contains the following steps:

Step 1- parsed tweets are stored after emptying lists
call to twitter API to get the tweets
parsing tweets one by one
saving text of tweet
saving sentiment of tweet
appending parsed tweet to tweets list

if tweet has retweets (In other words, reshared or replied to, make that it is added only once
parsed tweets returned
Step 2- Twitter_Client Class object is created
Step 3- Function call to get tweets

Table 2- Python queries to take the input from the user and search for tweets

```
query = input("What to search for? ")
tweets = api.get_tweets(query , count = 1000)
f = open('helloworld.txt','w')
```

3.4 Polarity determination

Creating a utility function to classify sentiment (positive, negative and neutral) of passed tweets using textblob's sentiment method

Table 3- Python query to classify sentiments based on the polarity

```
if analysis.sentiment.polarity > 0:
    return 'positive'
elif analysis.sentiment.polarity == 0:
    return 'neutral'
else:
    return 'negative'
```

3.5 Analysis

3.5.1 picking positive tweets from tweets and finding percentage of positive tweets

Table 4- Python query to pick positive tweets

```
ptweets = [tweet for tweet in tweets if tweet['senti
ment'] == 'positive']
```

3.5.2 picking negative tweets from tweets and finding percentage of negative tweets

Table 5- Python query to pick negative tweets

```
ntweets = [tweet for tweet in tweets if tweet['senti
ment'] == 'negative']
```

3.5.3 picking neutral tweets from tweets and finding percentage of neutral tweets

Table 6- Python query to pick neutral tweets

```
neutweets = [tweet for tweet in tweets if tweet['sen
timent'] == 'neutral']
```

3.5.4 printing positive, negative and neutral tweets (if necessary)
3.5.5 creating a pie-chart using the data obtained

To assess the opinions of the public to find out their point of view on the current scenario of IPL, a Google form was created and sent. The frequency of watching matches by the public was assessed. Along with this, a poll was conducted to check if the public enjoyed watching the matches at home or at the stadium. Positive and negative responses in relation with the no audience system, entertainment factor, etc. were also recorded.
Once 200 responses were received, Tableau was used to visualize the data. This tool was used to visualise the relation between the frequency of cricket watchers and the frequency of people going to the stadium. Along with this, sentiment analysis was done on the opinions received on the comments received in the form.

4. RESULT AND ANALYSIS

In this study, we conducted two analyses. The first one was with the live Twitter data, and the second one with a Google form containing specific attributes.

We analysed the sentiments of people from thousands of tweets after particular matches and categorized them intopositive, negative and neutral tweets. This process allows one to realizewhat the fans felt after their favourite team won or lost.

4.1 CSK vs MI

This was the first match held in the IPL 2020 league on 19th September, 2020. This match witnessed an audience of 20 crores which resulted in a record for the highest number of views for a sport. The match was won by CSK.Fig. 1represents the percentage of positive, negative and neutral tweets for this match.

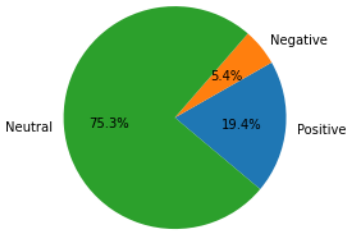

Figure 1- (CSK vs MI) tweet categorization

It is seen that most of the people (75.3%) expressed neutral tweets, that is they just expressed facts about the match devoid of any emotions.
A 19.4% of tweets expressed positive sentiments, that is people probably expressed their opinions saying that the teams played well.
A 5.4% of tweets expressed negative sentiments, that is people probably would have criticised the match or players

4.2 DC vs KXIP

This was the second match of the league which resulted in a super-over as the match was tied. It was a very interesting match to watch and it resulted in DC winning the match. Fig. 2 represents the percentage of positive, negative and neutral tweets for this match.

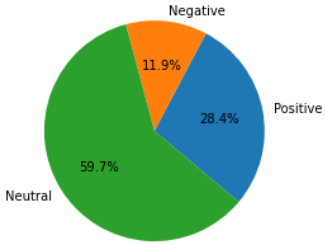

Figure 2- DC vs KXIP tweet categorization

4.3 MI vs RCB

This match was a very interesting one with RCB hitting a solid 201 runs and MI also hitting the same number of runs. This resulted in a nail-biting super-over. MI had hit 7 runs in the super-over and RCB finally won after achieving the required target. This match was definitely a treat to watch for all Cricket lovers.

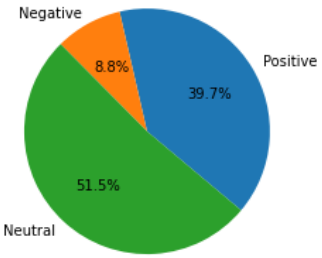

Figure 3- MI vs RCB tweet categorization

4.4 CSK vs RR

This match happened in the Sharjah stadium where a record was set for the highest number of sixes (33) during the match. RR set a high target of 217. CSK lost the match by 17 runs but the match was a delight to watch. Fig. 4 represents the percentage of positive, negative and neutral tweets for this match.

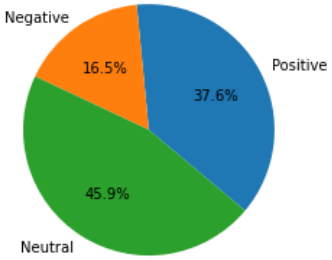

Figure 4- CSK vs RR tweet categorization

4.5 RR vs KXIP

This match happened in the Sharjah stadium again. KXIP hit a solid 223 runs in their first innings and RR successfully chased this target. This was the highest successful target chase in the history of IPL. It was a nail-biting match since the winner wasn't decided till the last ball. Fig. 5 represents the percentage of positive, negative and neutral tweets for this match.

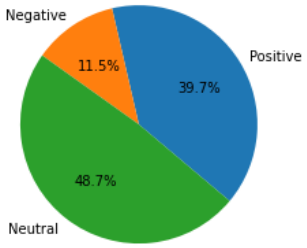

Figure 5- RR vs KXIP tweet categorization

Table 7- Table containing consolidated results of the sentiment analysis

Teams playing	% of positive tweets	% of negative tweets	% of neutral tweets
CSK and MI	19.4	5.4	75.2
DC and KXIP	28.4	11.9	59.7
MI and RCB	39.7	8.8	51.5
CSK and RR	37.6	16.5	45.9
RR and KXIP	39.8	11.5	48.7

Table 7 shows a consolidated result analysis of the five matches and the same is depicted in Fig 6. It is seen that there is a higher percentage of positive tweets than negative tweets. This indicates that the spirit of cricket does not decrease when a person's favourite team loses.

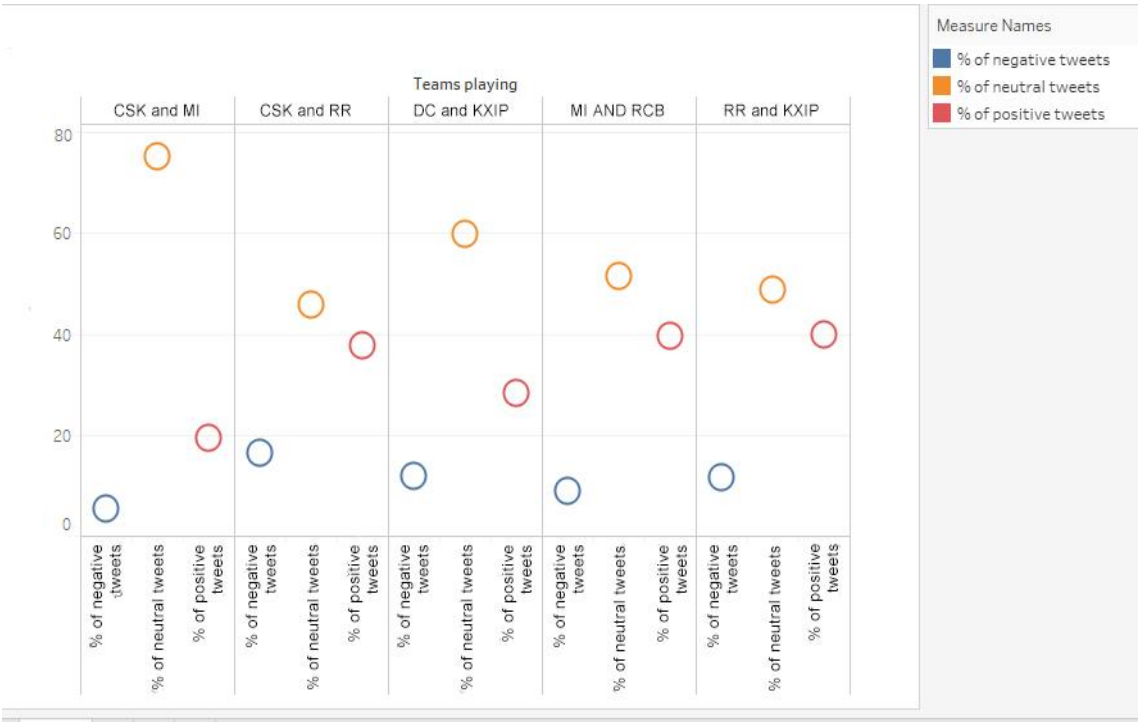

Figure 6: Comparison of Positive/Negative/Neutral tweets between teams

The second analysis was conducted using a Google form to understand the sentiments of people on the overall scenario of IPL happening during the Covid-19 situation. A whole lot of opinions were obtained. Fig. 7representsthe percentage of the polarity of those comments.

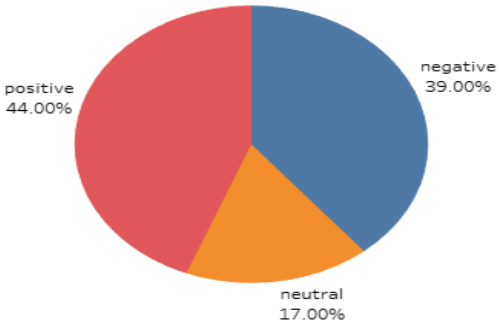

Figure7- Polarity checking on IPL during Covid-19

A few of the positive comments were “It’s good that they found a way to play even during corona time and I think it’s safe if they are being tested frequently.” and “It's nice that normalcy is returning with appropriate cautions in these tough times. Both the players and the fans have something to look forward to.”The neutral comments included “Amidst this hard time, it's good to see something entertaining that keeps us distracted from the crisis.” and “Missing the crowd and usual IPL atmosphere”.A few of the negative comments were “It lacks the same feel without the crowd at the stadium and all the usual vibe and build

up." and "It's a money minting project. In dire times, IPL wasn't necessary. What happened with CSK could have happened to other teams putting numerous at risk." (The comment relating to CSK was made because a few of the CSK players tested Corona positive.)

Overall, there was an almost equal percentage of people who had positive and negative views on IPL happening in Dubai amidst this entire pandemic situation.

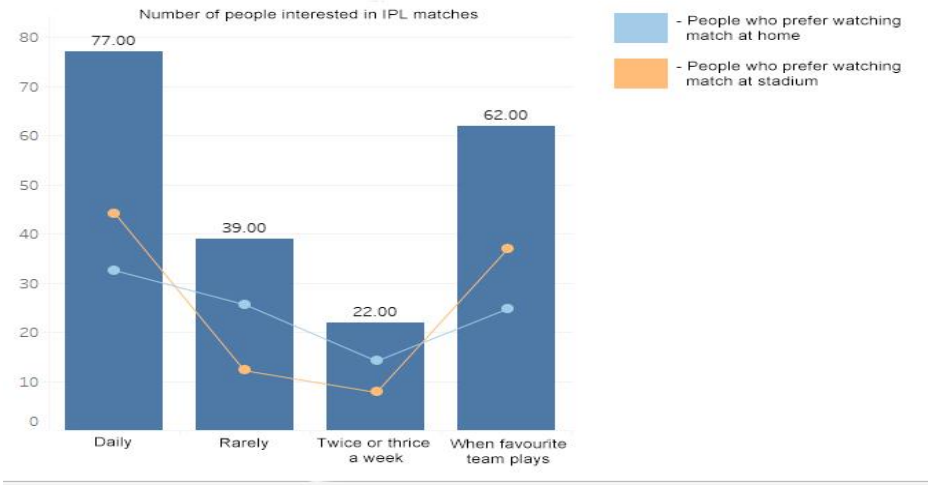

Figure8- Relation between frequency of watching match and venue preferred

Fig. 8shows us an analysis of the frequency of people watching an IPL match with the choice of watching a match at the stadium and watching a match at home. The yellowline shows the number of people who prefer watching a match at the stadium and the blueline shows the number of people who prefer watching a match at home. It is seen that the ones who daily watch matches prefer to watch a match at the stadium rather than watching it at their home. This shows the enthusiasm and love for cricket from the fans. Theones who watch matches rarely preferred to watch it at their home according to their comfort level.

5. CONCLUSION

IPL is not just another regular sporting event that occurs every few months in our country. It's a festival since it brings the entire country together. In this paper, we have presented the sentiment analysis for tweets after particular IPL matches. It was found that most of the tweets showed a positive sentiment. The data used is a random sample of live Twitter tweets containing particular keywords and the analysis was done using Python. Furthermore, we present an analysis on the sentiments of people on the topic of IPL during the pandemic situation. We got an almost equal percentage of positive and negative views along with a few neutral comments. The analysis suggests that most of them were willing to accept the situation but a few of them didn't want the event to start before the Covid-19 situation was resolved completely.

Although this year does stand out in contrast to the previous years in terms of the match experience since the fans are confined to their homes, the enthusiasm and energy has stayed the same, if not seen an evident spike! Watching their favourite teams and finding innovative methods to compensate the stadium cheer, IPL survives and has managed to bring a lot of smiles and laughter even in such difficult times.

References-

[1] I. Rish, An empirical study of the naive Bayes classifier, T J Watson Research Center, January 2001

[2] Mohammed J. Islam, Q. M. Jonathan Wu, Majid Ahmadi and Maher A. Sid-Ahmed, Investigating the Performance of Naïve- Bayes Classifiers and K- Nearest Neighbor Classifiers. Published in 2007 International Conference on Convergence Information Technology (ICCIT 2007)

[3] Bingwei Liu, Erik Blasch, Yu Chen, Dan Shen and Genshe Chen, Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier. Published in 2013 IEEE International Conference on Big Data

[4] Tan S., Cheng X., Wang Y., Xu H., Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis. In: Boughanem M., Berrut C., Mothe J., Soule-Dupuy C. (eds) Advances in Information Retrieval. ECIR 2009. Lecture Notes in Computer Science, vol 5478. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-00958-7_31

[5] Lina L. Dhande1 and Dr. Prof. Girish K. Patnaik, Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier. Volume 3, Issue 4 July-August 2014, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)

[6]BoyiXie, Apoorv Agarwal, Ilia Vovsha, Owen Rambow and Rebecca Passonneau, Sentiment Analysis of Twitter Data. Proceedings of the Workshop on Language in Social Media (LSM 2011), pages 30–38, Portland, Oregon, 23 June 2011.

[7] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts, Learning Word Vectors for Sentiment Analysis. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 142–150, Portland, Oregon, June 19-24, 2011.

[8] Alexander Pak and Patrick Paroubek, Twitter as a Corpus for Sentiment Analysis and Opinion Mining, 2010

[9] Theresa Wilson, Janyce Wiebe and Paul Hoffmann, Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 347–354, Vancouver, October 2005.

[10] Tetsuya Nasukawa and Jeonghee Yi Sentiment Analysis -- Capturing favorability using Natural Language Processing. Proceedings of the 2nd international conference on Knowledge capture, October 2003

[11] C. Troussas, M. Virvou, K. J. Espinosa, K. Llaguno and J. Caro, "Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning," *IISA 2013*, Piraeus, 2013, pp. 1-6, doi: 10.1109/IISA.2013.6623713.

[12] A. Go, R. Bhayani, L. Huang, “Twitter sentiment classification using distant supervision”, 2009

[13] PreslavNakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani and VeselinStoyanov, Sentiment Analysis in Twitter, arXiv:1912.01973v1 [cs.CL

[14] EfthymiosKouloumpis, Theresa Wilson, Johanna Moore, Twitter Sentiment Analysis: The Good the Bad and the OMG!, January 2011

[15] X Fang, X., Zhan, J. Sentiment analysis using product review data. Journal of Big Data 2, 5 (2015). https://doi.org/10.1186/s40537-015-0015-2