Sports Social Media Data Analysis and Classification

Nishad Tupe
Indiana University
Bloomington, IN, USA
ntupe@iu.edu

Sushant Athaley
Indiana University
Bloomington, IN, USA
sathaley@iu.edu

Izolda Fetko
Indiana University
Bloomington, IN, USA
ifetko@iu.edu

ABSTRACT

Twitter is one of the most vastly used social media platforms that allow individuals to voice their opinions on a broad range of topics, one of them is sports. Fans frequently use Twitter to show their admiration and support for their favorite teams and athletes, as well as dislikes towards the opponents. Our project is utilizing those types of tweets with the goal of analyzing sentiment around the cricket series between the West Indies and India team played in November 2018. In addition to the sentiment analysis, the project includes the use of the machine learning methods with the goal of classifying tweets based on team and individual player. The team had also explored the properties of player's interactions network present in the tweets and briefly test the hypothesis based on the network assortativity property. Our findings show that fan sentiment and their support varies significantly for a team and as well as for players. We also observed consistent performing players are discussed more often together. We believe our study can aid in the efforts to promote sports and may also help sports governing authorities to choose their ambassadors.

KEYWORDS

Twitter, Sports, Cricket, Sentiment Analysis, Classifier model, Data mining, Network analysis ,Hypothesis testing

1 INTRODUCTION

In the past decade, social media platforms have become an essential part of our society manifested in our daily interactions with the world. The usage of these platforms has grown immensely in recent years, making Twitter one of the most popular platforms for voicing one's opinions. According to Statista, Twitter has had approximately 335 million monthly-active-users (MAU) at the beginning of Q2 2018 [9]. This social media platform is often used to talk about sports events and competitions, among other topics. It allows sports fans to interact directly with celebrity athletes as well as with fellow fans.

Cricket is one of the favorite sports in the countries that experienced British colonization such as England, South Africa, India, Australia, Pakistan, and many more. According to some records from a few years ago, "it was estimated that over a billion people tuned into the Cricket World Cup game between India and Pakistan" [3]. The new shorter tournament formats like T20 also take much attention of the cricket fans, not just on TV but on social media sites such as Twitter. In our project, we observed the cricket enthusiasts never-short of expressing their feelings and showing their engagement into the game even for the cold series such as India and the West Indies, where one team is dominant over the other.

In this particular project, we focus on series in which the West Indies Cricket team tour India and played T20 format cricket matches. We analyze Twitter comments and interactions related to the #INDvsWI hashtag with the intent of conducting tweet classification and sentiment analysis. We chose this specific event due to its popularity and the fact that it is currently trending. We believe this tweet collection has given us an opportunity to learn how to use classification models appropriately and how to conduct a sentiment analysis, while at the same time allowing us to compare the results of our study to the actual outcomes of the competition.

In our study, we emphasize on sentiment analysis of social media users tweets. We strive to answer questions such as do both teams and its players have the same sentiment levels, how the sentiment vary as the series progress. In the process, we also build the classification model that predicts the tweets belongs to the player and team. We also make effort to understand the relationship between players that exists through cricket fan tweets and verify the nature of the network identified using the configuration model.

2 RELATED WORK

A significant amount of work concerning the sentiment analysis and classification of sports-related tweets has already been done. In their research, Apalak and Aparup Khatua analyze the Cricket World Cup 2015 Twitter data to test their model and "investigate the relationship between user classification (in a multi-team context) and user-level mix tweeting pattern" [6]. They used 3.5 million tweets collected during this tournament and similarly to our project employed a Logistic Regression model in the tweet classification phase [6].

In another study, a team of researchers had developed a system that collects data from Twitter and "performs sentiment analysis to determine whether they are positive, neutral, or negative and automatically visualizes them within multiple time-line representations" [4]. This system is called the Visual Twitter Analytics (Vista) and is made to "identify trends within the public sentiment in relation to one or more topics of interest as well as to support exploration and analytical reasoning" [4]. Although our project does not entail development of a new application such as Vista, the method used to conduct the sentiment analysis in our project is very similar to their approach as it includes classification of tweets into three buckets, positive, negative, and neutral.

Similarly to the "Vista" study, Zhao and the team built a webservice named SportSense that extracts television viewer sentiment related to sports from live tweets [12]. Their app is capable of recognizing major events in the National Football League (NFL) as promptly as 40 seconds into a game based on the analysis completed on the collection of those live tweets. The main question this study was trying to answer was the level of excitement of the game watchers and to find out "how positive their feelings are toward the game and each team of the game" [12]. This is in line with the goals of our project where we are looking to identify sentiment levels for both India and West Indies team as well as for individual player. Prior to their analysis, the data was pre-processed by removing links, re-tweets, emoticons, capital letters, question and exclamation marks. Once cleaned, the data was suitable for a lexicon-based approach to be completed. They "applied information retrieval techniques to generate a list of 20 most frequent words used over two million game-related tweets from 50 games played in 4 weeks of the 2010-2011 NFL season" [12]. This allowed them to identify parts of speech for each word (noun, verb, adjective, etc.) and produce a list of sentimental words. The team had found that among all tweets collected during those four weeks, 87% were positive, 11% were negative, while only 2% of tweets contained mixed sentiments [12].

Another work done by Pessoa and team included a sentiment analysis of the 2014 FIFA World Cup, a very popular soccer tournament that took place in Brazil [8]. In their paper, they analyzed the collective sentiments in the English language only, related to this sports event by completing the following steps: corpus composition, pre-processing, classification and temporal analysis [8]. The results of their classifier were binary (positive and negative sentiment), which also allowed them to compute the Collective Sentiment Score (CSS) in the temporal analysis stage[8].

A paper by Yu and Wang touches upon a similar topic. They took a big data approach and analyzed the U.S. soccer fan's tweet sentiments related to the 2014 FIFA World Cup as well [11]. They had found that "during the matches that the U.S. team played, fear and anger were the most common negative emotions and in general, increased when the opposing team scored and decreased when the U.S. team scored" [11]. On the other hand, more joyful emotions and feelings of anticipation were recorded more when the U.S. team would score and during matches between other teams when the pressure was equal to zero for the U.S. team [11].

A paper published by Megha Arora and co-authors on the Indian Premier League (IPL) depicts the correlation between the star-studded teams and their popularity on various social media platforms such as Facebook and Twitter [2]. In their study, they analyzed 2.6 million tweets and utilized sentiment analysis to identify emotion levels of the IPL fans related to spot-fixing allegations [2]. The similarities between Aurora's project and this analysis is the fact that both studies used the sentiment analysis as the primary method to explore the tweets.

Based on our research, we firmly believe that information collected from social media platforms such as Twitter (in this case for sporting events) goes beyond the sentiment analysis and can be re-purposed for various implementations such as the promotion of sports, aiding sports franchises, player contracts, and even finding athlete advocates that can positively influence individuals and larger public masses as well.

3 DATA AND METHODS

We followed the step by step approach as shown in the process diagram (Figure 1) to analyze and publish the results of the study.

We collected the data from Twitter for the event #INDvsWI from 3rd of November to 11th November. Once collected, the tweets were cleaned and labeled which made them ready for the sentiment analysis and classification. Upon this step, the team conducted an exploratory analysis with the purpose of getting familiar with the data. The next step involved applying the sentiment analysis which helped us understand the emotions/sentiments around each team and player. After completing sentiment analysis, the team moved forward with completing the tweet classification based on team and player tweets. In the following step, the team had created a network graph based on the name mentions for each player in the tweets, to conclude our network analysis. In the final step of this process, the team published the results and conclusion of their study.

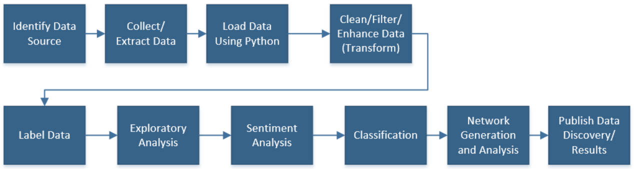


Figure 1: Methodology

3.1 Dataset

The Twitter platform provides a tagging option for each event and players, which becomes very helpful when trying to find relevant data for a particular analysis. Due to those convenient options, we decided to use Twitter as the primary source of our study. We used the Twitter public API to retrieve the tweets involving the India vs. West Indies cricket series using the #INDvsWI hashtag. The team concentrated on the T20 series which was a three-match series played between these two nations in the time frame between November 3, 2018, and November 11, 2018. For this time frame, the team collected approximately 27 thousand tweets. The final match of the series played on November 11th and the tweets of the final day used for classification purpose. During the data collection step, the team had captured the following tweet attributes: tweet time, tweet text, and tweet source. The tweet time feature used for various time-series analysis. tweet text was used to conduct the sentiment analysis, classification analysis, and network identification.

3.2 Data Labeling

Data labeling is an essential step for any supervised machine learning activity. This step is known to be more complicated when working with social media data, as text labeling can be tricky due to its dependence on the context of a particular text. In this study, we intended to classify tweets based on the team which it belongs to, as well as the players participating in the event. There are typically manual and automated are the two methods that can be applied to social media data labeling. In the manual approach, one can manually go through each tweet, and based on the understood context classifies/labels the text. This method can be very accurate concerning labeling. However, it can be very time consuming and usually generates less data for the classifier. On the contrary automated method finds a way to label the data accordingly pragmatically. This method helps generate a more significant amount of labeled data, however, may not be as accurate as the manual method.

We labeled data for two categories in this study, one label generated for team name and another label for player name based on tweet text. This study uses automatic labeling as we could find out the team name and player name in the tweet and accordingly we can generate the labels. This method provided a sizable amount of labeled data for the machine learning classifier.

We generated team label by searching team name in every tweet to understand tweet is about which team. It is possible that tweet may contain team names multiple times or no team name at all. We applied weight factor by counting the number of time team name present in the tweet. Based on the higher count for the team name, the tweet is labeled for that particular team either India or West Indies. In case, no team name present we applied a dummy label so that we can ignore that data from the classification task. This method provided us with approx 7k labeled tweet for the team name. Similarly, we generated labels for the players by searching player name in the tweets. List of all players from both teams participated in the tournament was prepared. We captured all variations of the player name like first name, last name and twitter account handle and created a dictionary so that we can search for all possible variation in the tweet. We then counted the number of occurrences of each player name in the tweet and the player with the highest count is labeled for that particular tweet. We could label approx 14k tweets for a player name using this method.

This labeled data then used during sentiment analysis and text classification analysis to find insight by team and players.

3.3 Sports tweets Classification

3.3.1 Team and Player Classification : In the team & player-based classification, we built a models to classify tweets based on a player and team information. We used Scikit-learn logistic regression classifier.

3.3.2 Sentiment Classification Lexicon Method : In the sentiments analysis, we build the model that classifies the tweet has a positive or negative sentiment using different Lexicon and Machine learning methods. In the Lexicon based approach, we used TextBlob and Affin libraries. The TextBlob library is based on the NLTK and boasts many advantages like sentiment analysis, pos-tagging, noun phrase extraction, etc. [5] AFINN is one of lexicon relatively easy and can extend to use for sentiment analysis. "The current version of the lexicon is AFINN-en-165. txt and it contains over 3,300+ words with a polarity score associated with each word. The author has also created a nice wrapper library on top of this in Python called afinn" [7]. Using the polarity scores, we then marked the tweet with classes such as positive(score above zero), negative(below zero) else neutral. Once we have labeled dataset we then performed exploratory visualizations to see the sentiment trends in the tweets for both team and players using polarity score and polarity classes.

3.3.3 Sentiment Classification Machine Learning Method : In real life classification problems are more prevalent than regression. The classification methods of machine learning algorithm allow us to predict the results of unknown data. In the machine learning-based sentiment analysis, we used scikit-learn's logistic regression and random forest classifiers to predict the positive and negative sentiment of the tweets.

3.4 Interaction Network

tweeter provides a capability to tag players or any tweet account holder in a tweet. Usually, people tag players, friends or celebrities in tweets, depending on the context of those tweets. This method of tagging provides an opportunity to understand how people are tagged and how they are related. Network analysis of such tagging can provide a good insight of most discussed players and how they are related. It can provide perspicacity into various player clusters or communities of the players. Community formation also helps to understand the context in cases where a particular player is outside of a network or not associated with any other player. Network analysis is an excellent way to understand how players interact with each other during particular sports events.

Generally tagging in the tweet done by using account name of the user which starts with @ symbol. In a particular tweet, any number of accounts can be tagged. We developed this network by extracting all account mentioned in a tweet and associating those accounts with each other. This association provided a relationship between various accounts which is considered as an edge in the network. Nodes in the network represent the user accounts. The network we developed was consist of 440 nodes accounts and 8062 edges.

3.5 Network Configuration Model

In the graph, we observed star players like Rohit Sharma and Virat Kohli strongly connected and having higher node degrees. This observation encouraged us to verify the hypothesis around assortative mixing property. In network science, "assortativity is a preference for a network's nodes to attach to others that are similar in some way " [10]. "The assortativity coefficient is the Pearson correlation coefficient of degree between pairs of linked nodes" [10]. As such r=1 then the network is said to be completely assortative, high degree nodes connect to only high degree nodes. r=0 network is considered to be nonassortative while r<0 network is said to be completely disassortative, that mean node with a higher degree also tend to connect lower degree nodes [10]. In network science, we can use the null models for hypothesis testing. Null models often considered for determining whether certain graph features are responsible for some characteristic of the graph or some pattern of behavior on the graph. To create the null model that describes the actual graph properties, we can use configuration models that allow us to hold the degree sequence constant while investigating other graph characteristics After calculating the degree assortativity, we then compared our player's graph to the null model to see if all graph properties are explained simply by the assortativity itself. Then we calculate the Z scores such that large Z score rejects the null hypothesis and confirm that we cannot just consider the degree sequence as feature is not sufficient to explain the entire properties of network and there could more special properties [1]. We used python's networkx library to load and calculate the graph properties such as assortativity and clustering etc.

4 RESULTS

In this section, we will discuss results and observation from our study.

4.1 Data Exploration

During the exploratory analysis, we found that the majority of tweets were submitted by the team India supporters, more specifically 75.2%, which can be seen in Figure 2. A significantly smaller percentage of tweets attributed to the Windies team, totalling in 22.8%. Based on the analysis, only 2% of tweets were undetermined, in other words, were not able to be classified as one or other team.

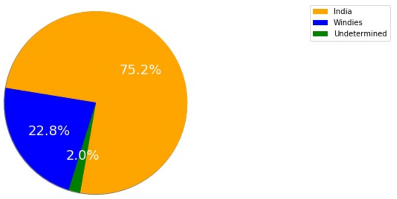


Figure 2: tweets by Team

We also explored the types of sources from which the tweets were made public as shown in Figure 3. The definite winner of the top ten sources was the Twitter for Android, with more than 17,800 tweets, followed Twitter Web Client and Twitter for IPhone which totalled in a little over 2,200 tweets each. IFTT, Twitter Lite came in fourth and fifth with more than a 1,000 tweets during this sports event, while other sources with less than this number of tweets included the tweetDeck, Cloudhopper, Mobile Web (M2), Facebook, and CricketNDTVLive.

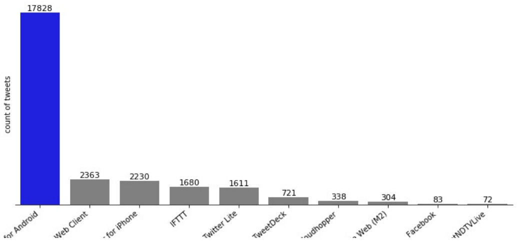


Figure 3: Top Ten Sources

When looking at the data by hour, we observed majority of tweets was tweeted in the hours between 12pm and 19pm, peaking at 16pm for both the Indian and the Windies team. As expected, the number of tweets was pretty low in the hours before noon, and would quickly increase in the afternoons as the time of the game was approaching. This trend was noticeable in all three categories for the two teams, as well as the group of undetermined tweets.

4.2 Sentiment Analysis

The following sections shows the results for lexicon based and machine learning based methods.

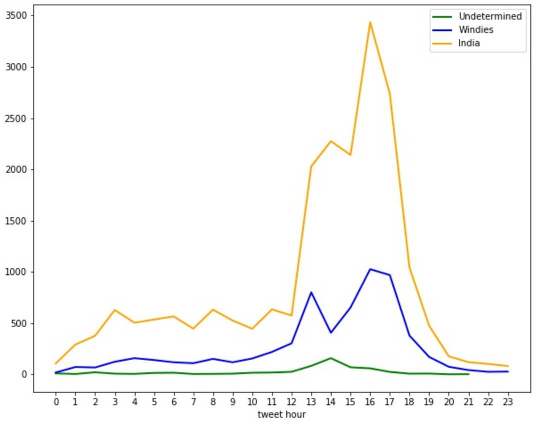


Figure 4: tweets by Hour

4.2.1 *Lexicon Method.*

From the exploratory visualizations of the sentiment scores, the data have shown us that the players of the Indian team favored more than the West Indies team players. Although the Indian team is considered to be very strong and features world-class players, during this event, we observed new players such as Khaleel , Krunal Pandya received a lot of fan attention. Interestingly, legendary players such as Virat Kohli and MS Dhoni were still among the top tweeted players despite not participating in series shows their legacy. The figure 5 shows top tweeted players with along their sentiment scores. Overall the sentiments results are very much in line with the actual performances by the players in the series. Players such as Rohit Sharma, Shikar Dhawan, Dinesh Kartik, Khaleel showed excellent performances.

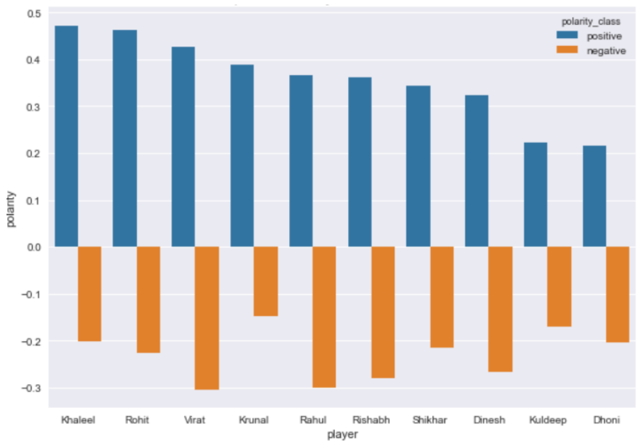


Figure 5: Top Players Sentiment Score

As shown in figure 6 team India was found to be crowd's favorite, while it proved to be a cold series for the West Indies team as they were playing away. With the fact that top-ranked team India played on their home soil undoubtedly resulted in a more significant positive vibe from their Twitter fans. Sentiment score graph thus explains that the two teams had different sentiment levels and the team India is Twitter community's favorite team.

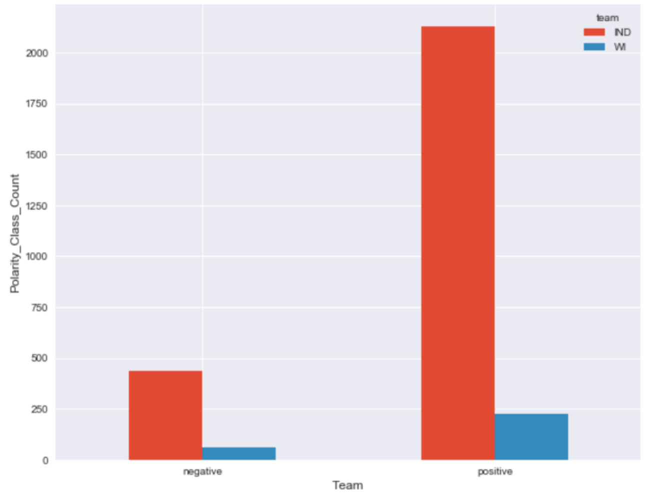


Figure 6: Team Sentiment Classification

The following figure shows the overall pscore values for both teams over the days as the series progressed as we can see again from the violin plot the team India had more sentiments throughout while sentiments for West Indies faded just before the end of the series.

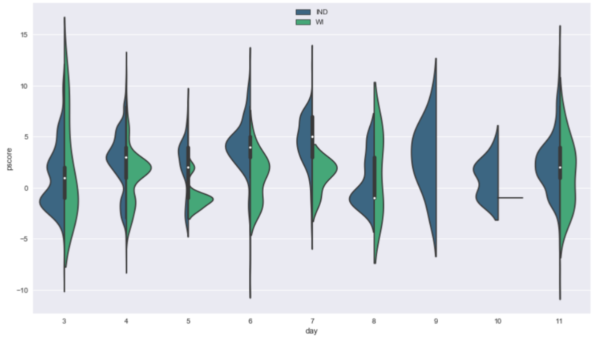


Figure 7: Pscores by Team

4.2.2 Machine Learning Method. We used Random Forest and Logistic Regression algorithms to perform the classification of the tweets. Before applying the machine learning methods, we performed the standard steps such cleaning the data, Extract the features, Labeling the data, etc. As discussed previously, we labeled our tweets as positive negative based on scores received from Lexicon Libraries Textblob and Afinn.

Table1 describes the Random Forest, Logistic Regression model results to classify the tweets as positive or negative.

Table 1: Sentiment Analysis ML Algorithm Scores

Algorithm	Accuracy Score
Random Forest	93%
Logistic Regression	88%

4.3 Classification

One of the primary goals of this study was to understand if we can classify tweets for team/player and to predict most talked about the entity on the final day of the match. We divided our tweets on a train and test dataset by considering the final match data as a testing set and the rest of the days as a training set. We used a logistic regression algorithm for the classification and applied a ten-fold cross-validation attribute.

The linear logistic regression model was trained on tweets in such manner to classify the data based on the team name - India or West Indies and showed the accuracy score of 98.60%. This trained model was later used to predict the tweets of final days match. The prediction model yielded 95% of tweets for team India and 4% of tweets for the West Indies team. The final T20 match played on November 11th, 2018 won by team India. Our model had proved that team India was the most talked about team on the final day of the tournament, which is in line with the results of the actual tournament.

The team also trained a multinomial logistic regression model in order to classify tweets based on the player's name. This model's accuracy was 93%. This model was then used to predict the final day's tweets. Using this model, the most talked about player predicted were Rohit Sharma, Dinesh, Rishabh Pant, Shikhar Dhawan, and Manish. All these players belonged to team India and had played well during the final day's match, which proved that our model had classified the final day's tweets correctly.

Table2 shows accuracy for our machine learning model.

Table 2: Classification Model Accuracy

Classification	Accuracy
Team	98%
Players	93%

4.4 Network

We also wanted to study network aspect of the sporting event, so we generated network based on players or tweeter users in the tweets. We established connections between players based on our philosophy that if two or more players talked about in particular tweet than most likely those are related to each other. We then visualize this network using python network library and Gephi tool. Gephi tool provides us with an excellent network visualization. Figure 8 shows the interaction network of the players.

This network is consists of 6 significant communities driven by the players. Nodes which are denoted using @ are either player's name or twitter user's name. The strongest node which is @imro45(Rohit Sharma) was captaining team India for this cricket tournament and is the center of this network which means he is discussed a lot with other players in this tournament. Rohit has a strong connection

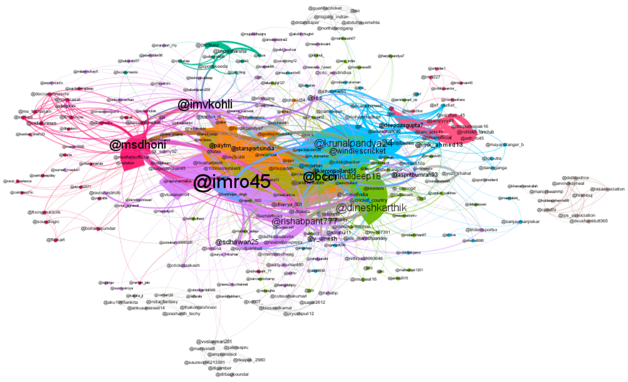


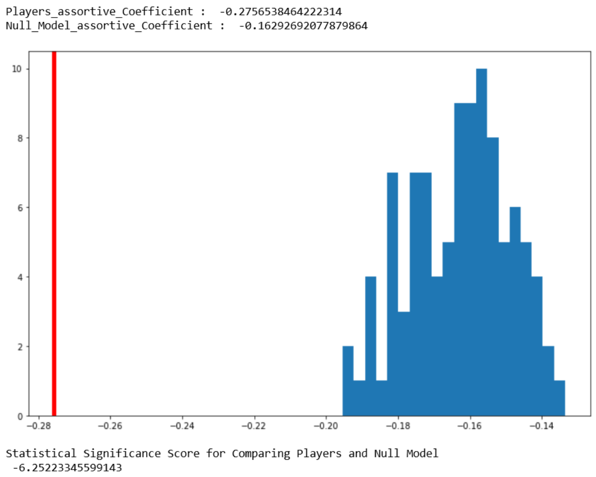
Figure 8: Interaction Network

with other Indian players @msdhoni(MS Dhoni), @imvkholi(Virat Kohli), @krunalpandya24 (Krunal Pandya), @rishabpant777 (Rishab Pant), @dineshkartik(Dinesh Kartik), @sdhawan25 (Shikhar Dhawan) and @bcci(Board of Control for Cricket in India). We observed the community in pink centered around MS Dhoni which is a former India cricket team captain. Community in blue color centered around Krunal Pandya and community green color centered around Dinesh Kartik. It is interesting that people discussed MS Dhoni and Virat Kholi although they were not playing in this particular series. It implies that players get discussed in social media even if they are not playing may be for the comparisons or other reasons. There is a strong bond between MS Dhoni and Rishab Pant which make sense as both players play the same role of the wicketkeeper, and their comparison is obvious. There are some interesting nodes like @paytm, @starsportsindia, @bcci which are not the players but institutions which are discussed by tweet users along with the players. We also observed that there is less mention of West Indies players which may be due to the poor performance or small fan base or due to they were not playing in their home country.

This network gives us an idea of how people worship their player and how communities are formed around the players. It also highlights important players of the sporting event along with their contribution as they are denoted with bigger nodes. We can also figure out which players are connected and how strong bond is between those players, this can be very useful as team bonding is the foundation for any team game. This network also highlights shortcomings for a particular team as well, for example, there are very few West Indies players present in the network which provides an improvement opportunity to find out how this sport can be promoted in that country for more prominent adaptation. This network is useful to understand the bond between the players and find out the improvement opportunities to strengthen that bond.

4.5 Hypothesis Testing

We got a degree assortativity value of -0.27 suggets the network is having somewhat of disassortative properties. As we can see from the graph, larger nodes such as Rohit Sharma and Virat Kohli, MS Dhoni also connect to lower nodes, the lower negative score suggests that the network is not entirely disassortative also. In either case, it was important to compare players graph assortativity with the null model to check if there is anything of interest to be pursued here. To verify our hypothesis statistically, we calculated the Z scores of assortativity coefficients of null models and real graph. As shown 9 the distribution of assortativity of the Null model. A red line in the figure shows players network assortativity With Z-score of -6.25 which not close to zero, we can safely conclude that just the degree sequence of the existing network does not explain the entire properties of players graph. With our hypthesis though we do observe that performing players are tagged along more often than others however they are also mentioned or tagged with lower degree nodes at times and have their own fan base.



```
Players_assortive_Coefficient :  -0.2756538464222314
Null_Model_assortive_Coefficient :  -0.16292692077879864
```

```
Statistical Significance Score for Comparing Players and Null Model
 -6.25223345599143
```

Figure 9: Networks Assortativity

5 CONCLUSION

The modern technology today allows those digitized emotions to be analyzed, classified, grouped into communities and predicted in various ways. In this paper, we have proposed methods to analyze a data collected from Twitter, classify the tweets, and sentiment analysis using machine learning and lexicons. We largely analyzed the sentiment levels of the tweets and observed that team India and its players were much more favored than the significantly weaker team of West Indies. We also build the models that performed fairly accurately whether to classify tweets based on team and player or classifying the sentiments associated with it. In the end, we were able to extract the complex relationship that coexists in social media data through users tweets. This network is driven and revolved around the star or performing players and help identify important players. We also show areas where the network is weak or not available and that can help with an opportunity to promote the sport. Our configuration model showed that star players not just mentioned together often on social media but also have their own fan base which can visualize through communities seen in the network. We believe our study methodology can be applied to any sporting event to understand not only high points but also the improvement opportunity to keep alive this wonderful event of Sport.

6 LIMITATIONS AND FUTURE WORK

We used automated labeling in this study for based on the team and players names which do not consider the context of the text. This automated labeling gave us a significant amount of data to our machine learning classifier, but we can think of a more robust method to perform the labeling instead of just counting the name frequency. We used regex to clean the data by using commonly used patterns however instead of using manual methods overall the data cleaning process could be improved by using tools or additional methods. The lexicon we used does not detect all type of sentiments especially sarcasm and also may not give importance to the emoticons. In the future, we expect to extend our study by overcoming the highlighted challenges to have a better predictive model.

ACKNOWLEDGMENTS

The author would like to thank professor Ali Ghazinejad for his support and suggestions in this project.

7 WORK CONTRIBUTIONS

Contributors listed in alphabetical order for each task.

- Project Methodology - Sushant Athaley, Izolda Fetko, Nishad Tupe
- Data Gathering - Sushant Athaley, Izolda Fetko, Nishad Tupe
- Data Cleaning - Nishad Tupe
- Data Labeling - Sushant Athaley, Nishad Tupe
- Abstract, Introduction, Literature review - Izolda Fetko
- Data Exploration Visualization - Izolda Fetko
- Data Exploration Analysis - Izolda Fetko
- Sentiment Model and Analysis - Nishad Tupe
- Classification Model and Analysis - Sushant Athaley
- Network Generation - Sushant Athaley
- Network Visualization - Sushant Athaley
- Network Analysis - Sushant Athaley
- Network Configuration Model - Nishad Tupe
- Network Hypothesis - Nishad Tupe
- Conclusion - Sushant Athaley , Izolda Fetko, Nishad Tupe
- Project Papers - Sushant Athaley, Izolda Fetko, Nishad Tupe

REFERENCES

[1] Y Y Ahn. 2018. Network Science Spring -2018, Lecture Notes:Applying the configuration model. Github - Jupyter Notebook. (Feb. 2018). https://github.com/yy/netsci-course/blob/master/m09-randomgraphs/configuration_model_assignment.ipynb

[2] Megha Arora, Raghav Gupta, and Ponnurangam Kumaraguru. 2014. Indian Premier League (IPL), Cricket, Online Social Media. Web Page. (may 2014). https://arxiv.org/pdf/1405.5009.pdf

[3] The Guardian. 2015. Cricket World Cup: India v Pakistan watched by a billion people in pictures. Web Page. (feb 2015). https://www.theguardian.com/sport/gallery/2015/feb/15/cricket-world-cup-india-v-pakistan-watched-by-a-billion-people-in-pictures

[4] Orland Hoeber, Larena Hoeber, Laura Wood, Ryan Snelgrove, Isabella Hugel, and Dayne Wagner. 2015. Visual Twitter Analytics: Exploring Fan and Organizer Sentiment During Le Tour de France. Web Page. (2015). http://www2.cs.uregina.ca/~hoeber/download/2013-vis-sdv.pdf

[5] Shubham Jain. 2018. Natural Language Processing for Beginners: Using TextBlob. Web Page. (Feb. 2018). https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-textblob/

[6] Apalak Khatua and Aparup Khatua. 2017. Cricket World Cup 2015: Predicting User's Orientation Through Mix Tweets on Twitter Platform. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (ASONAM '17)*. ACM, New York, NY, USA, 948–951. https://doi.org/10.1145/3110025.3119398

[7] Himanshu Lohiya. 2018. Sentiment Analysis with AFINN Lexicon. Web Page. (July 2018). https://medium.com/@himanshu_23732/sentiment-analysis-with-afinn-lexicon-930533dfe75b

[8] Rubens Pessoa, Jonathas Magalhaes, Marlos Silva, Henrique Pacca, and Evandro Costa. 2015. 2014 FIFA World Cup: An Initial Analysis of Collective Sentiments in Twitter. Web Page. (2015). https://pdfs.semanticscholar.org/efa5/61c4ae578b2a010f3d23dabafbb65960f34f.pdf

[9] Statista. 2018. Number of Monthly Active Twitter Users Worldwide from 1st Quarter 2010 to 2nd Quarter 2018 (in millions). web. (July 2018). https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[10] Wikipedia. 2018. Web Page. (Sept. 2018). https://en.wikipedia.org/wiki/Assortativity

[11] Yang Yu and Xiao Wang. 2015. World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fans' tweets. 48 (07 2015). https://dl.acm.org/citation.cfm?id=2781900.2782141

[12] Zhong L. Wickramasuriya J. Vasudevan V. Zhao, S. 2011. Analyzing Twitter for Social TV: Sentiment Extraction for Sports. Web Page. (2011). https://pdfs.semanticscholar.org/d99c/21adb61a5d080e4de3b52a4d1c01fc4719e0.pdf