Twitter Sentiment Analysis: A Survey in Cricket and Bollywood

Nayantara Kotoky*, Smiti Singhal†, Anushka Sharma‡ and Dhara Ajudia§
*Applied Neurocognitive Systems, Fraunhofer Institute for Industrial Engineering, Germany
†‡§Department of Computer Engineering, Pandit Deendayal Energy University
Gandhinagar, India
Email: *nayantara.kotoky@gmail.com, †smitis2000@gmail.com, ‡anushkas0706@gmail.com, §dhara.ajudia1108@gmail.com

***Abstract*—Twitter has been the voice of the public for a long time now. With the rise in the usage of Twitter, the active participation of its users in expressing their views across all domains has significantly increased. This paper aims to perform sentiment analysis and study the influence of Bollywood and Cricket celebrities on Twitter users. Three different types of information are extracted from the tweets using sentiment analysis, namely, (1) sentiments of people towards cricket, cinema (Bollywood), and gender, (2) identifying the highly discussed individuals and events for each category, and (3) co-occurrence analysis for identifying closely discussed celebrities belonging to different categories. Our analysis identifies that females in the cricket sport are not as popular as compared to their male counterparts whereas females in the entertainment industry (Bollywood) are equally popular as the males. We also identify current trends that are the target of discussion in Twitter using Network of Words analysis. In addition, the co-occurrence analysis shows very high association between Male Cricketers and female Bollywood stars. In essence, we try to determine the emotional tone of people to gain an insight of the hidden attitudes and opinions expressed in a tweet regarding cricket and Bollywood.**

***Keywords*—Twitter; Sentiment analysis; Text mining; Co-occurrence Network.**

I. INTRODUCTION

Twitter is a micro-blogging site that has become a public forum for anyone who wants their voice to be heard. People post their views in the form of short messages consisting of words, images, or videos, up to 280 words, and these posts are famously known as tweets. The recent numbers suggest that over the years, the Twitter market has seen notable growth in its number of users in all major developing or already developed countries. Hence, big companies and brands try to understand the sentiments of people through the variety of views from a plethora of tweets. These sentiments are studied in order to get meaningful results that help them understand the collective opinion on their services.

Twitter has been widely used for performing sentiment analysis on various topics like politics [1] [10], entertainment [11], etc. A large number of researchers have studied Twitter sentiment analysis on the basis of a number of factors such as polarity, or sentiments like anxiety, anger, etc. Several tools have been created for identifying various sentiments, some of which include Linguistic Inquiry and Word Count (LIWC) [1], TextBlob [8], Valence Aware Dictionary for Sentiment Reasoning (VADER) [5] [10], and Orange [2].

In this paper, we have collected data from Twitter for two categories - Entertainment and Cricket, to perform sentiment analysis and classify them on the basis of the emotions attached to those tweets. The objective of this research is to understand the influence of cricket sport and Bollywood cinema, two very popular sources of entertainment in India, and understand how they can be utilized as media to influence the masses and bring societal changes. Our work analyzes public sentiments associated with individuals in these two groups as well as a collective outlook into cricket and Bollywood. With the insights drawn from several analyses on tweets regarding these topics, we identify a few ways in which people's emotions are impacted.

We have collected a total of 6000 tweets using Tweepy which included 20 hashtags for every category. The collected tweets are tagged as Cricket, Bollywood, male and female depending on the subject of the tweet. Using the tweets, we have identified the sentiments of people regarding the four categories *cricket-male, cricket-female, bollywood-male, bollywood-female* and then recognized the celebrities in Cricket and Bollywood which are two groups that are very influential among the people in India. Furthermore, an in-depth analysis of co-occurrence in the four different categories was performed to determine the association between these categories and to investigate which categories are mentioned together and why.

The main contributions of this research work are:

- Sentiment analysis is performed on the tweets using VADER and Tweet Profiler. The results show that the tweets consisting of the happiest emotion of the people are shown for the category of women actresses and women cricketers.
- KH Coder [16] (named after the developer Koichi Higuchi) is used to carry out performance analysis of the celebrities in the four categories using their corresponding tweets where they are mentioned. The analysis uncovers tweets in support and criticism of certain people showing close correspondence to real-life phenomena.
- Co-occurrence analysis of the four categories of tweets is performed using Jaccard Coefficient. The results show that *bollywood-male* and *bollywood-female* are the two categories that are highly associated with Twitter discussions.

 ISBN: 978-1-61208-994-2

Outline: The paper is structured as follows. Section 2 gives insights into the research done in this domain and the various papers published in this area. Section 3 briefly describes the objective and implementation details and guides through the approach. Next, Section 4 covers in detail the results obtained and the interpretations of the analysis. Section 5 concludes the paper.

II. Related Work

In this section, we mention specific articles which are used for this research purpose. The literature contains twitter analysis performed for understanding people's emotions for various domains and their use. This section also discusses certain tools and their utility in performing sentiment analysis.

A. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment

Tumasjan et al. [1] determine whether Twitter can be used to predict the result of the federal election of the national parliament in Germany. 104300 political tweets were first translated from German to English and the text analysis software "Linguistic Inquiry and Word Count (LIWC)" was used to get the results. In conclusion, the results depicted that Twitter can be undoubtedly regarded as a plausible indicator of political opinion.

B. Content Analysis of Dark Net

Nattuthurai and Aryal [9] used KH Coder to analyze the data from the darknet using Co-occurrence network analysis. The data collected was categorized into business-related and non-business-related data. Multi-Dimensional Scaling and Co-occurrence network analysis were performed on the dataset to uncover a higher frequency of negative words associated with both categories.

C. Analysis of Data Using Data Mining tool Orange

Kukasvadiya et al. [2] discuss the concept of data mining and how the Data Mining tool Orange performs when subjected to any kind of data. The paper provides a practical implementation of Orange. It concludes how Orange is easier than others and can perform a wide range of data analysis using its widgets such as sentiment analysis, visualizing Time series data and plotting heatmaps.

Marcu et al. [12] analyzed data related to educational aspects collected from various high schools using Orange and classified them according to Ekman and Plutchik models of emotions. Tweet Profiler, a feature provided by Orange, classified the data based on these models and both the results were compared to find the best solution for sentiment analysis.

Thange et al. [13] have worked upon a COVID-19 dataset of India and have visually represented the relationships in the dataset using Orange. As a result of their analysis, it has been found that there have been more cases of infection in men compared to women and maximum number of infected patients are in the 30 years age-group.

D. KH Coder: An exploratory analysis of the text mining of news articles about "water and society"

Hori [3] aims to discover social interest in the issues of water and society from media reports and to compare it in Japanese and international media. This research uses the online databases of two newspapers: the Japan News and the International New York Times. The social interest is discovered by cluster analysis, that is, to derive clusters that have value with respect to the problem being addressed. The articles extracted from those databases are analyzed using KH Coder and the generated co-occurrence network.

E. Twitter Sentiment Analysis Using Natural Language Toolkit (NLTK) and VADER

Elbagir et al. [10] aim to compare two powerful sentiment analysis tools - NLTK and VADER on the data collected for the 2016 US presidential elections from the microblogging service Twitter. The analysis concludes how VADER was an effective and better choice for sentiment analysis classification.

III. Proposed Work and Implementation Details

The purpose of the paper is to analyze the tweets and compare the influence, joint mentions, and other different aspects of the text to get meaningful results. The three main experimental analysis are as follows:

1) Classification of sentiments
2) Twitter as a reflection of performance
3) Relative frequency of joint mentions inter-category and intra-category

A. Data Collection

Around 6000 tweets were collected as part of the dataset. The time span considered is January 2021 to November 2021. It is to be noted that the collected tweets span for the specific time frame and for the specific categories of interest, that is, Indian cricket and Bollywood (Indian cinema). The methodology used for data collection is as follows [4]:

- Importing Tweepy - an easy-to-use Python library for accessing the Twitter API
- Authentication for Twitter Developer account
- Defining list of hashtags for every category
- Defining the *date_since* date as a variable
- Filtering retweets
- Output data as .csv file

With the use of Twitter API, recent tweets for most popular persons of all 4 categories were collected and merged together. For instance, #ViratKohli, #RohitSharma for *cricket-male*, #SmritiMandhana, #HarmanpreetKaur for *cricket-female*, #RanbirKapoor, #KartikAaryan for *bollywood-male*, #AliaBhatt, #DeepikaPadukone for *bollywood-female*, to name a few hashtags considered. The distribution of the data for each category can be seen in Figure 2.

In order to verify the actuality of the data collected, document map was used. Document map, a feature in the Orange Tool for text mining, shows geolocation from the textual data (here, tweets). It finds the mentions of countries/capitals

 ISBN: 978-1-61208-994-2

(whenever it is present in a tweet) and displays the frequency of occurrence in the world map. Around 125 tweets mentioned the name of a country/capital which is used to create the document map.

In Figure 1, we can see that the number of mentions of India (red) is considerably higher than the other countries and these countries have been accurately displayed because of the 2021 T20 World Cup (T20 is an international cricket world cup tournament which consists of 16 teams competing with each other in a twenty-over cricket match. India was one of the top 12 teams to play in the 2021 T20 world cup and thus was vigorously discussed on Twitter). The participants in the T20 World cup were Namibia, Pakistan, Afghanistan, England, Bangladesh, Australia, etc. which are highlighted on the world map too.

B. Classification of sentiments

The tools we used for sentiment analysis were TextBlob [8], Empath, Pattern, Sentiwordnet and VADER. The most precise results were shown by VADER [5], which accurately categorized the sentiments into positive and negative.

C. Twitter as a reflection of performance

Orange tool [6] was used for the sentiment analysis of tweets based on the seven basic emotions proposed by Ekman [14]: fear, anger, joy, sadness, disgust, appreciation, and surprise.

D. Relative frequency of joint mentions inter-category and intra-category

KH Coder is a text mining tool which is typically used for finding the potential relationships between entities represented within a document [7]. Here, we first load the required file as the data and then pre-process it as a necessity for analysis. The tool is implemented to obtain co-occurrence matrices and networks to get insights into major themes from the text and analyze the associations between the text that appear together.

IV. Results and Interpretations

A. Classification of sentiments

1) Visualizations between categories: Figure 2 shows the amount of positive and negative emotions for all the four categories. From the figure, we observe that:

- The positive sentiments for the almost same amount of tweets are the highest for Female Bollywood Stars.
- The negative sentiments reflected through the tweets are the highest for Male Cricketers.

This suggests that people publicly post their opinion about celebrities and, even though Male Cricketers are the most famous personalities, as we see them more frequently in other social media discussions and advertisements, it does not discourage people from openly sharing negative views about them.

B. Twitter as a reflection of performance

1) Results for Male Cricketers: Figure 3 shows the polarity of sentiments of the tweets for individual Male Cricketers. We observe the following:

Highly Appreciated: MS Dhoni

In light of the 2021 T20 World Cup, people were excited and happy because of Dhoni's presence as the mentor for Team India.

Involved in Negative Discussion: Axar Patel and Krunal Pandya

Board of Control for Cricket in India (BCCI) [15] had announced through their social media handles that Axar Patel would be replaced by Shardul Thakur in the T20 matches. This did not go well with a lot of people and hence people shared their disappointment through Twitter. Negative emotions were shared using Axar Patel as the topic, although the negativity was not necessarily targeted toward him. On the other hand, the negative emotion towards Krunal Pandya is due to his poor performance in the Indian Premier League (IPL) matches that made the Twitterati furious. Here, we see negative emotions being expressed but under two different circumstances, the first one supporting the player and the second one being targeted toward the player.

2) Results for Male Bollywood Stars: Figure 4, visualization for Male Bollywood Stars, shows the following results:

Highly Appreciated : Ranbir Kapoor

Despite any recent project announcements or any other controversies, Ranbir Kapoor still remains the most highly appreciated actor because of his huge fan following.

Involved in Negative Discussion: Nawazuddin Siddiqui

Nawazuddin Siddiqui's statement garnered attention due to its critical comment on racism being a bigger issue in the Bollywood Industry as compared to nepotism. People supported him and highly criticized the Bollywood Industry.

3) Results for Female Celebrities (Cricketers and Bollywood Stars): On a similar pattern, the other 2 categories showed the following results.

As shown in Figure 5, Yastika Bhatia, under Female Cricketers category, received huge appreciation for her brilliant innings in one of the league matches. Also, it can be clearly seen that no female cricketer received harsh criticism.

Among female Bollywood Stars, as per Figure 6, Ananya Pandey being highly active on social media was amongst the favourites while news of Nora Fatehi being involved in money laundering led to unfavourable discussions.

4) Sentiment Analysis using TweetProfiler: Tweet Profiler, a widget provided by orange, retrieves information about the emotions attached to the sentiment by sending data to the server where a model calculates the emotion scores/probabilities according to the text. This is then plotted with the help of a Box Plot. This analysis provides seven different sentiments in comparison to only positive and negative sentiments.

Sentiment Analysis of Female Cricketers using Tweet Profiler is shown in Figure 7. The figure clearly explains that Yastika Bhatia and Priya Punia have the highest percentage

 ISBN: 978-1-61208-994-2

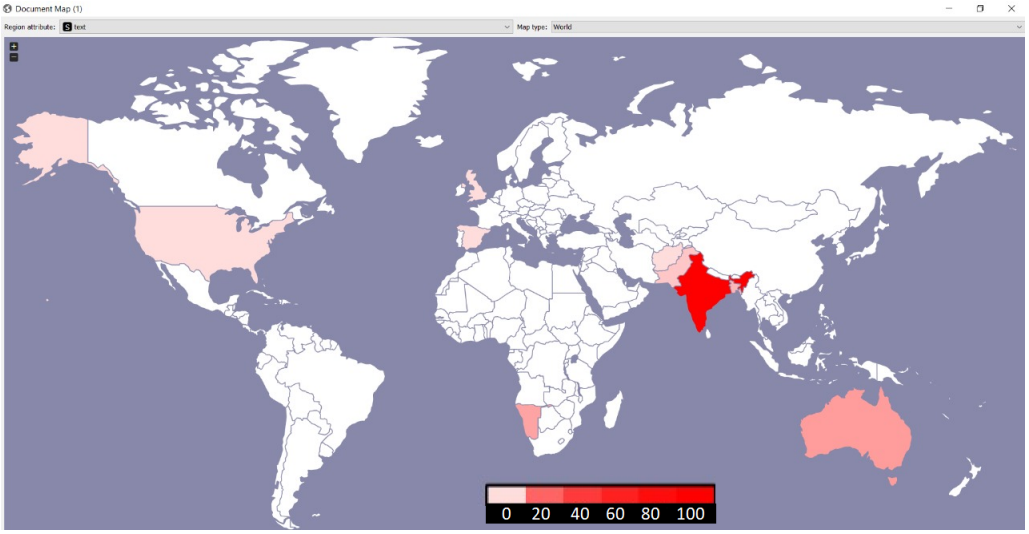


Fig. 1. Document Map for Male Cricketers.

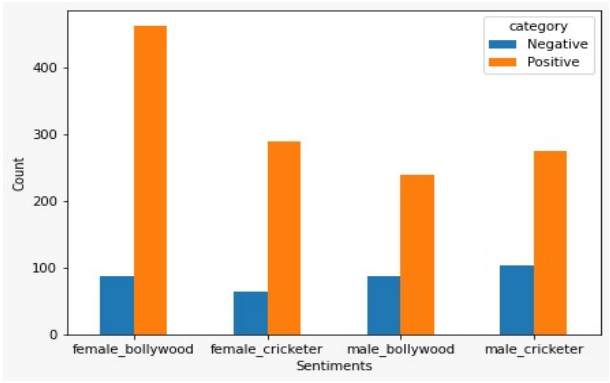


Fig. 2. Visualizations between categories.

of joyful tweets while the tweets mentioning Shikha Pandey depicted highest percentage of surprise emotion.

Yastika Bhatia and Priya Punia are indicated as having the highest positive sentiments using VADER (Result IV-B3, Figure 5) as well as the most joyful emotions using Orange (Figure 7).

C. Relative frequency of joint mentions inter-category and intra-category

This analysis clusters the nodes that have similar occurrence. In this work, it means that nodes which are often mentioned together by people's tweets are clustered, where nodes represent individuals from the four categories. The representation shows the words with similar appearance patterns, that is, with high degrees of co-occurrence, connected by lines. To determine edge strength, Jaccard coefficients are calculated for all possible combinations of target words. It was carried out on a combined document for all categories. Figure 8 shows the network of words showing clusters of closely discussed individuals. From these clusters, we can identify real-life events that led to the discussion. The following are the interpretations for Figure 8:

1) Network of Words:

- Green (03) - Shows the discussions regarding the announcement of the film "Vikram Vedha" starring Hrithik Roshan and Saif Ali Khan.
- Light Orange (12) - This cluster depicts the discussions on the recent film Sooryavanshi, and it can be concluded that through these tweets we can correctly make out the major cast and director of this film.
- Blue (01) - Talks on the T20 World Cup 2021.
- Orange (02) - This cluster comprises the tweets related to KKR Vs DC semi-finals and KKR vs CSK finals having the common factor, KKR in IPL.

2) Network of Codes: This analysis plots a network diagram to explore the association of people of different categories.

Male and Female Cricketers: Figure 9 shows the association between male and Female Cricketers. The network shows how rarely the Female Cricketers are mentioned along with the Male Cricketers. There is a clear separation of categories of Female Cricketers yellow (02) and orange (06) and the link with the Male Cricketers green (01) and blue (05) is extremely weak. Additionally, the size of the circles also shows that the frequency with which the Male Cricketers are discussed with each other within the category is much greater than when they are mentioned with the Female Cricketers. Also, the frequency with which the Female Cricketers are mentioned together within their category is much lower in frequency (which can be deduced by the size of the circle), indicating that Female Cricketers do not receive as much popularity among the Twitter people as Male Cricketers.

Male and Female Bollywood stars: Figure 10 shows the network of codes for the Bollywood fraternity. This analysis shows an intricate connection between the males and the females which brings us to the conclusion that these categories

 ISBN: 978-1-61208-994-2

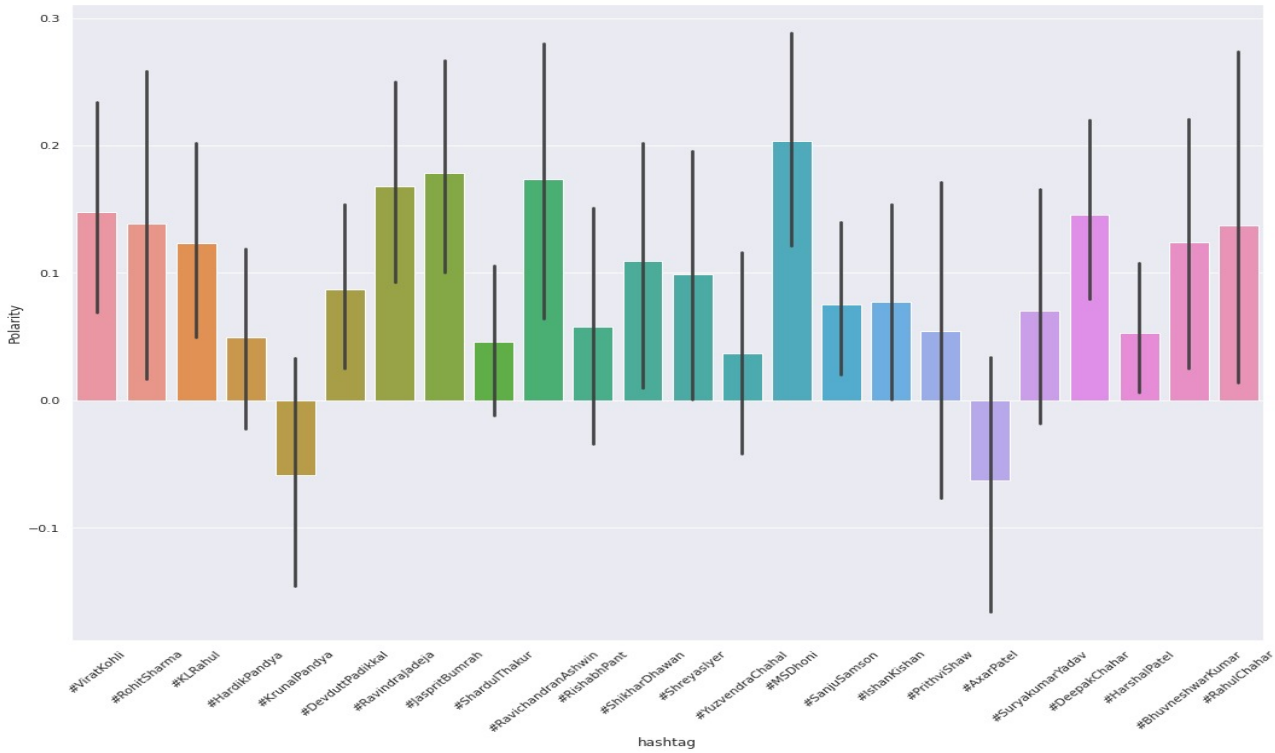


Fig. 3. Visualization of positive and negative sentiments for Male Cricketers using VADER.

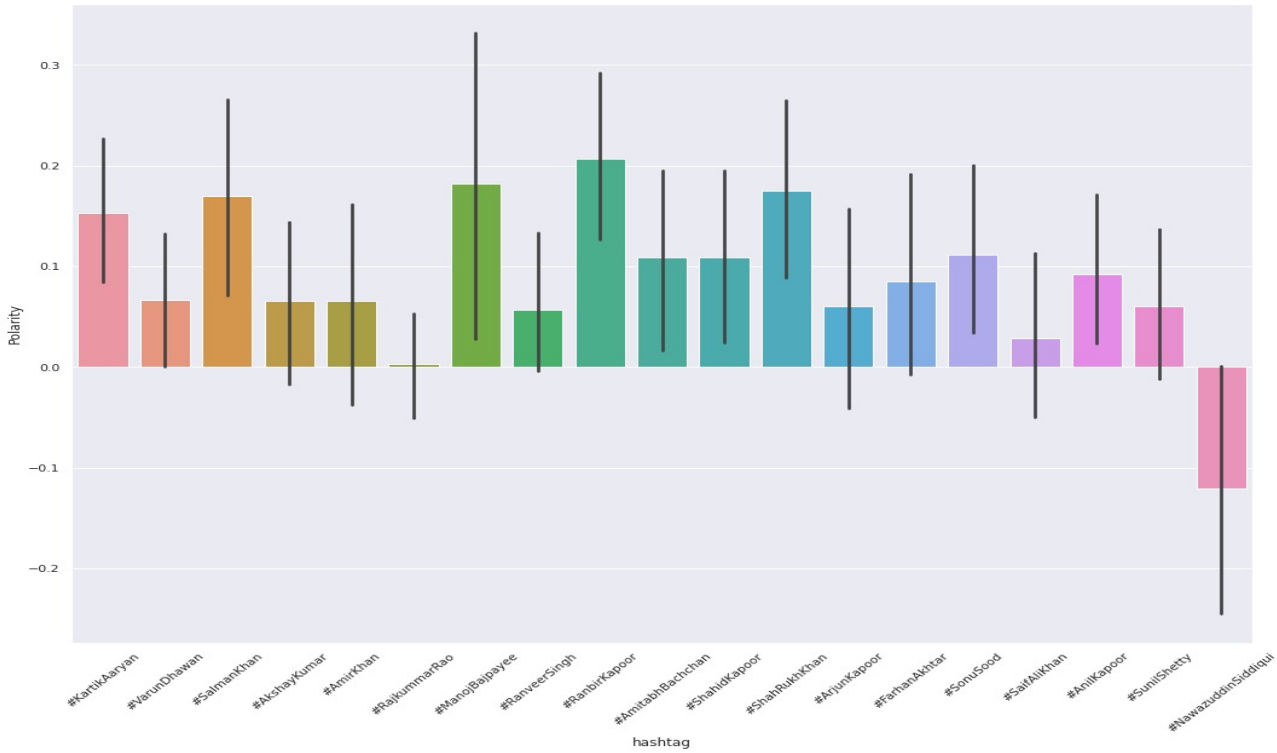


Fig. 4. Visualization of positive and negative sentiments for Male Bollywood stars using VADER.

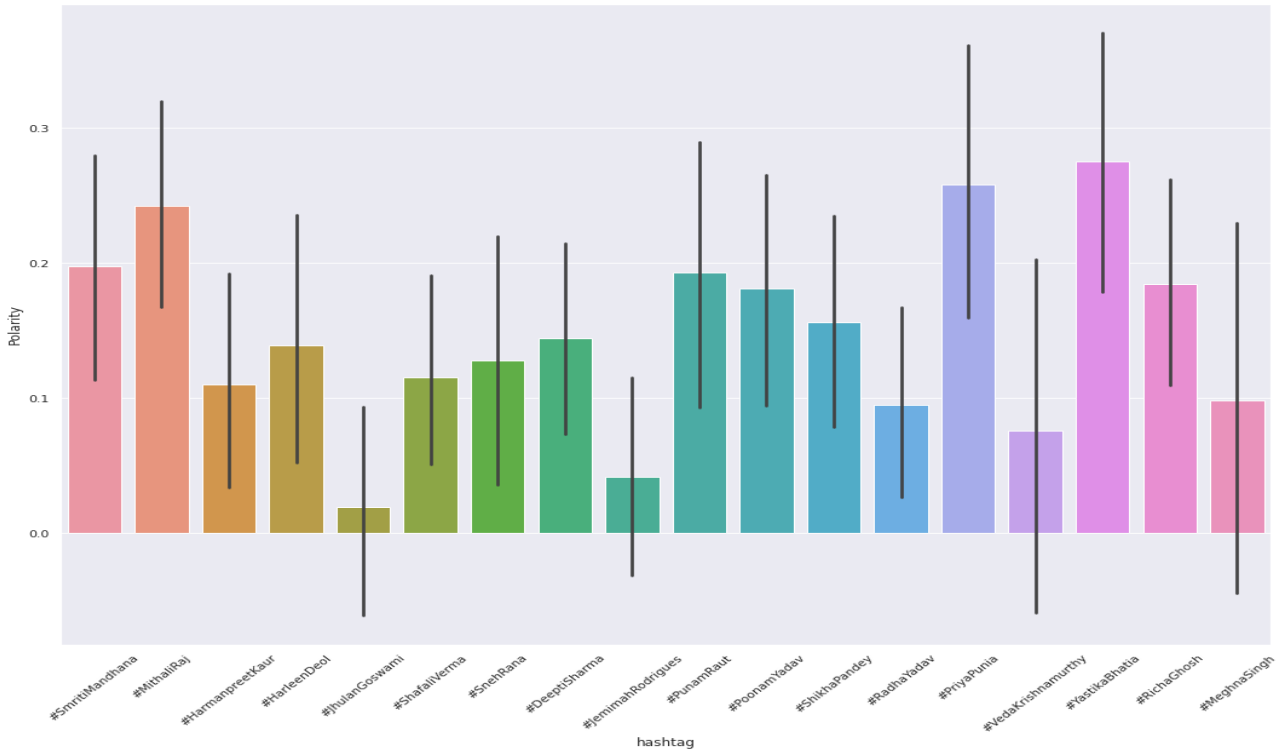


Fig. 5. Visualization of positive and negative sentiments for Female Cricketers using VADER.

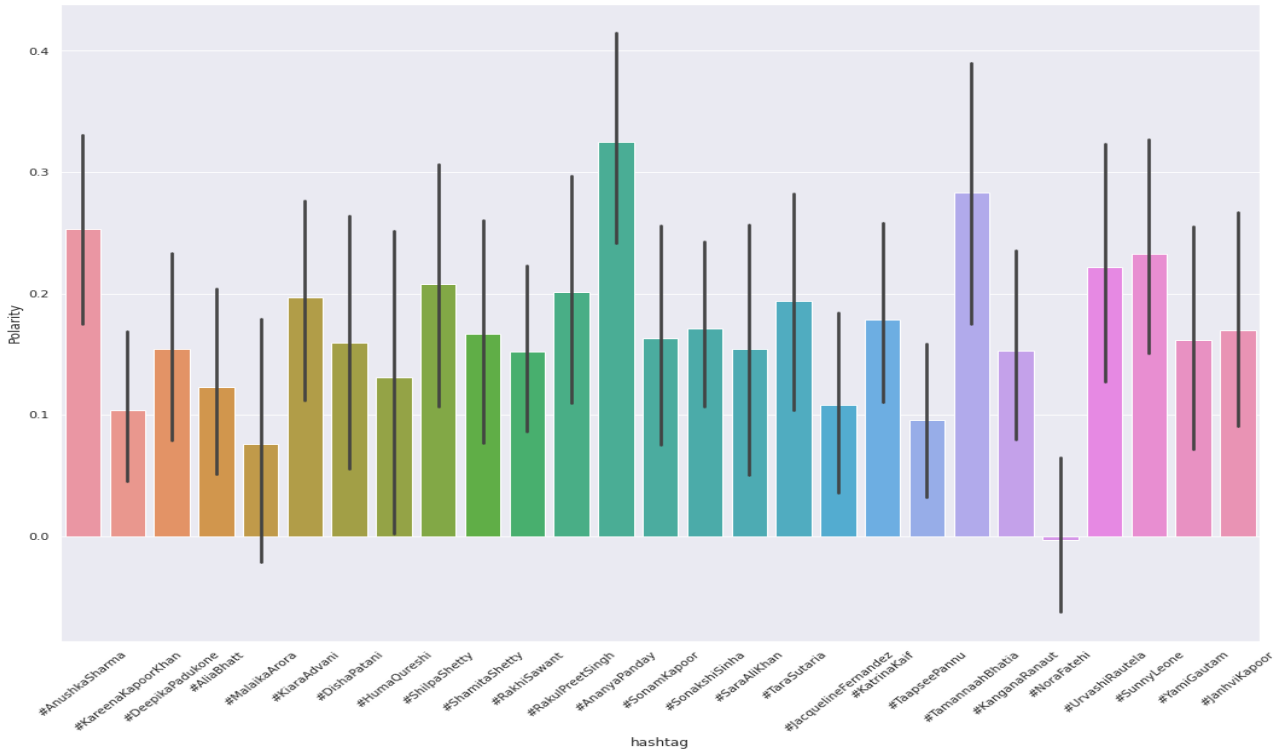


Fig. 6. Visualization of positive and negative sentiments for Female Bollywood Stars using VADER.

 ISBN: 978-1-61208-994-2

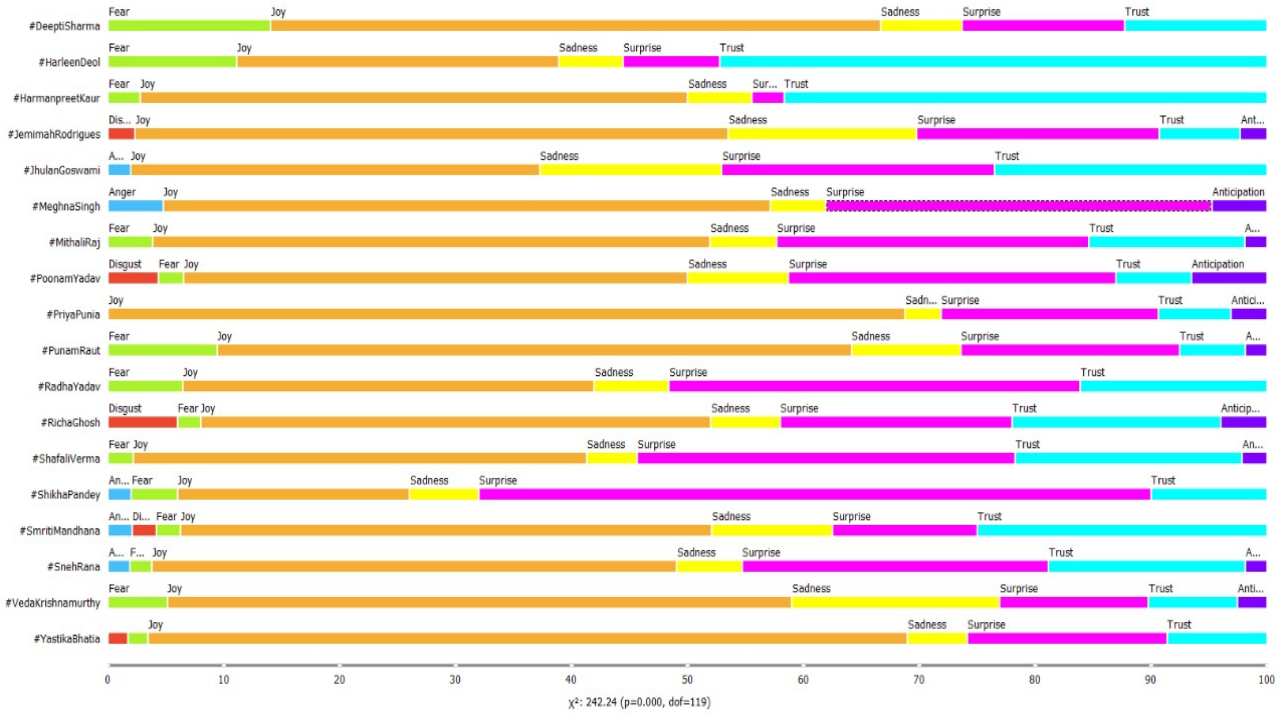


Fig. 7. Sentiment Analysis of Female Cricketers using Tweet Profiler.

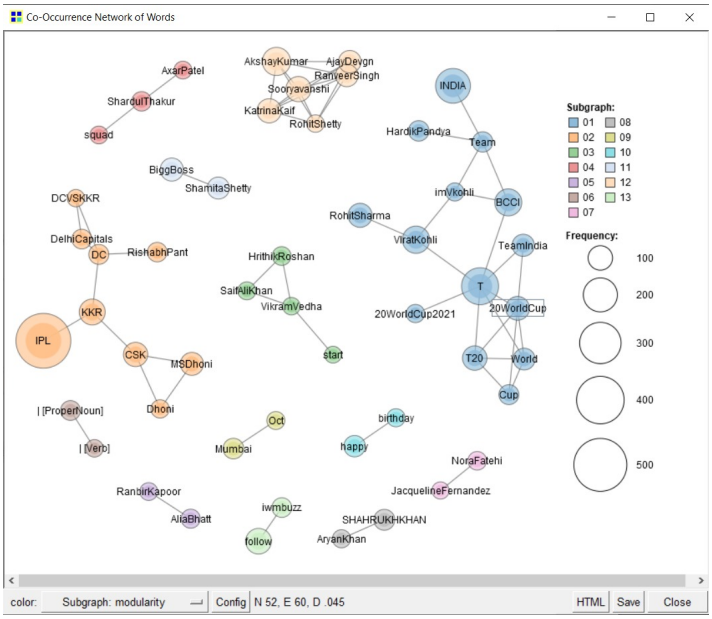


Fig. 8. Network of Words for all categories combined.

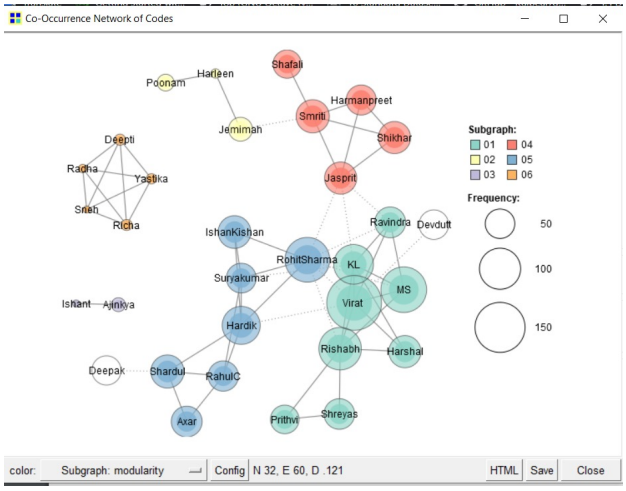


Fig. 9. Network of Codes for Male and Female Cricketers.

are often mentioned together and are not under-shadowed by each other. Unlike the case of the cricket sport, the entertainment industry enjoys similar popularity and discussion among the Twitter people for both males and females. For example, the green cluster (01) consists of the cast of Sooryavanshi, and Ranbir (male) and Alia (female) are mentioned together in many tweets.

3) Co-occurrence Matrix: Figure 11 shows the Jaccard coefficient that shows the association of two individuals with respect to how much they are discussed together in the tweets. In Figure 11, the red circle depicts the Jaccard coefficient for Virat Kohli (male cricketer) and Smriti Mandhana (female cricketer).

The co-efficient is calculated as follows:

$\frac{|Virat \bigcap Smriti|}{|Virat \bigcup Smriti|} = 0.004$

(The number of times both were mentioned together divided

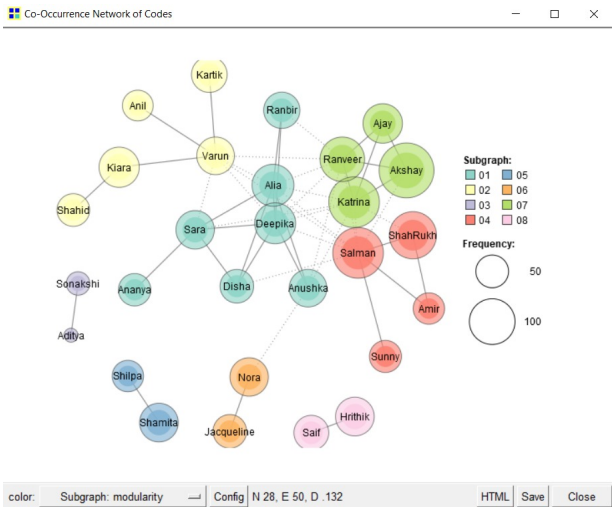
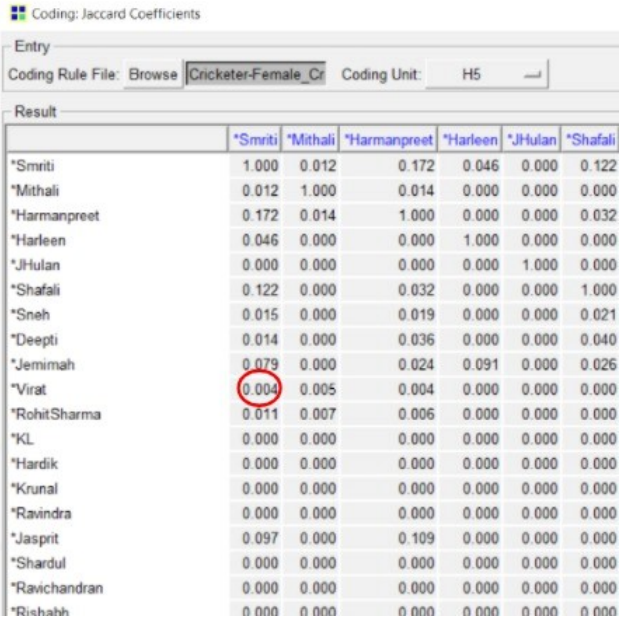


Fig. 10. Network of Codes for Male and Female Bollywood stars.



	*Smriti	*Mithali	*Harmanpreet	*Harleen	*JHulan	*Shafali
*Smriti	1.000	0.012	0.172	0.046	0.000	0.122
*Mithali	0.012	1.000	0.014	0.000	0.000	0.000
*Harmanpreet	0.172	0.014	1.000	0.000	0.000	0.032
*Harleen	0.046	0.000	0.000	1.000	0.000	0.000
*JHulan	0.000	0.000	0.000	0.000	1.000	0.000
*Shafali	0.122	0.000	0.032	0.000	0.000	1.000
*Sneh	0.015	0.000	0.019	0.000	0.000	0.021
*Deepti	0.014	0.000	0.036	0.000	0.000	0.040
*Jemimah	0.079	0.000	0.024	0.091	0.000	0.026
*Virat	0.004	0.005	0.004	0.000	0.000	0.000
*RohitSharma	0.011	0.007	0.006	0.000	0.000	0.000
*KL	0.000	0.000	0.000	0.000	0.000	0.000
*Hardik	0.000	0.000	0.000	0.000	0.000	0.000
*Krunal	0.000	0.000	0.000	0.000	0.000	0.000
*Ravindra	0.000	0.000	0.000	0.000	0.000	0.000
*Jasprit	0.097	0.000	0.109	0.000	0.000	0.000
*Shardul	0.000	0.000	0.000	0.000	0.000	0.000
*Ravichandran	0.000	0.000	0.000	0.000	0.000	0.000
*Rishabh	0.000	0.000	0.000	0.000	0.000	0.000

Fig. 11. Co-Occurence Matrix.

by the number of times either of them was mentioned)

We counted the results, which lie between 0 and 1 (exclusive), to get the count of the number of cases where the celebrities were mentioned together for all cases and they were combined to get the final results. The final count for the co-occurrence matrix is shown in Figure 12. We observe:

- While the maximum joint mentions are of Male Cricketers with fellow Male Cricketers (count=233), the joint mentions of Female Cricketers with female actors (1) and male actors (2) are negligible.
- It can be concluded that Female Cricketers, in general, were tweeted much less with other men as well as women compared to Male Cricketers. The popularity of Female Cricketers is less among the Indian Twitter people.

	Actor Female	Actor Male	Cricketer Female	Cricketer Male
Actor Female	163	176	1	17
Actor Male	176	136	2	18
Cricketer Female	1	2	54	15
Cricketer Male	17	18	15	233

Fig. 12. Co-Occurrence Matrix Result.

- It is observed that there are quite a high number of tweets mentioning male and female actors together (count=176). This analysis shows resonance to the real-life phenomenon where movies include both genders (unlike cricket teams), and hence their count of joint mentions is much higher than the other category.

V. Conclusion

In this paper, we present three different analyses on how discussions are held regarding Indian actors and cricketers using Twitter as a platform of expressing opinions. Here, Sentiment Analysis was performed mainly using 3 tools - VADER, Orange, and KH Coder on a dataset of around 6000 tweets, collected with the help of Twitter API Tweepy. With three distinct analysis methods, we have identified several interpretations of how the Twitter people view the four categories, that is, Male Cricketers, Female Cricketers, male actors and female actors.

Interpretations were drawn based on the outputs obtained from the experiments. Some significant observations were the less mention of women cricketers compared to other categories (Section IV-C2) and the noteworthy association between Bollywood (Females and Males) and Cricket (Males) Figure 12. Also, analyzing the most appreciated and criticized celebrities helps the brands publicize their products to customers by connecting with those celebrities for marketing purposes. This eventually helps the brands in making their product famous by attracting customers. Furthermore, we also observed that a well-balanced overview of current affairs can be acquired by looking at the significant amount of tweets in light of the latest happenings. This leads us to believe that Twitter can be seen as a reliable platform to view the actual sentiments of the people given our current analysis and context.

The findings of this study strengthen the fact that Twitter is an effective platform for resonating with the audience. Twitter Sentiment Analysis lets users ascertain the vibe of a conversation and gives leverage to users as they are able to delve deeper into the emotions involved in interactions. Although the analysis is performed at the Indian context, the methodology is generic and can be extended to other countries or world-wide topics. In addition, the analysis techniques can be used for identifying people's emotions on various other topics like war, usage of specific technology, natural phenomenon like climate change, etc.

 ISBN: 978-1-61208-994-2

REFERENCES

[1] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," The International AAAI Conference on Web and Social Media 16 vol.4, No.1, May 2010, pp. 178-185, doi:10.1609/icwsm.v4i1.14009.

[2] M. Kukasvadiya and N. Divecha, "Analysis of data using data mining tool orange," International Journal of Engineering Development and Research 5.2, June 2017, pp. 1836-1840, ISSN - 2321-9939.

[3] S. Hori, "An exploratory analysis of the text mining of news articles about water and society," WIT Transactions on The Built Environment, 168, 2015, pp. 501-508, doi:10.2495/SD150441.

[4] Automate Getting Twitter Data in Python Using Tweepy and API Access [Online], Available from: https://www.earthdatascience.org/courses/use-data-open-source-python/intro-to-apis/twitter-data-in-python/

[5] A. Beri, "Sentimental Analysis Using Vader, interpretation and classification of emotions". [Online] Available from: Aditya Beri, https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664 [retrieved: November, 2022].

[6] Orange widget catalog, https://orangedatamining.com/widget-catalog/text-mining/twitter-widget/ [retrieved: November, 2022].

[7] KH Coder 3 Reference Manual, Available from: https://khcoder.net/en/manual_en_v3.pdf[retrieved: November, 2022].

[8] Sentiment Analysis using TextBlob. Parthvi Shah. Available From: https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524.

[9] P. Nattuthurai, and A. Aryal, "Content Analysis of Dark Net: Academic Journals from 2010-2017 Using KH Coder," ACET Journal of Computer Education and Research 11, 2018, pp. 25-35.

[10] S. Elbagir and J. Yang, "Twitter sentiment analysis using natural language toolkit and VADER sentiment." Proceedings of the international multiconference of engineers and computer scientists. Vol. 122, 2019, pp. 16.

[11] Y. Yu and X. Wang, "World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fans' tweets, Computers in Human Behavior", Volume 48, 2015, pp. 392-400, ISSN 0747-5632, doi:https://doi.org/10.1016/j.chb.2015.01.075.

[12] D. Marcu and M. Danubianu, "Sentiment Analysis from Students' Feedback : A Romanian High School Case Study," 2020 International Conference on Development and Application Systems (DAS), 2020, pp. 204-209, doi:10.1109/DAS49615.2020.9108927.

[13] U. Thange, V. Shukla, R. Punhani and W. Grobbelaar, "Analyzing COVID-19 Dataset through Data Mining Tool Orange," 2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM), 2021, pp. 198-203, doi:10.1109/ICCAKM50778.2021.9357754.

[14] P. Ekman, Basic emotions Handbook of cognition and emotion, 1999, pp. 16.

[15] BCCI Official Website. [Online]. Available from: https://www.bcci.tv/

[16] KH Coder. [Online]. Available from: http://khcoder.net/en/

 ISBN: 978-1-61208-994-2