Blockbuster: Predicting movie success using social network community sentiment analysis

Sandeep Ranjan[1], Dr.Sumesh Sood[2]

[1]*PhD Scholar, I.K.Gujral Punjab Technical University, Kapurthala (India)*
[2]*Assistant Professor, I.K. Gujral Punjab Technical University, Kapurthala (India)*

ABSTRACT

Market research has been significantly affected both positively and negatively by the recent developments in social media.Movie promoters try to predict the movie performance on the box-office through various methods.Social networks like Twitter have recently been studied for their role in accurately predicting the success of movies. The experiment proposes a mathematical model based on the sentiment analysis of the community tweets mined for movie Twitter hashtags. For this research experiment, Twitter was minedto create a dataset oftweets for the Twitter handles of Bollywood movies released between 1stJanuary 2017 and31stMay 2017. Online word of mouth was detected as constituent communities of the dataset. Sentiment analysis was performed on all the 19,055 tweets by assigning polarityweights to individual tweets. The proposed prediction model used overall sentiment score of movies to predict their success on the box-office and the results were statistically validated using the actual box-office verdict. The model results achieved 0.833correlation with the actual verdict.

Keywords: -box-office success, graphs, online word of mouth, opinion mining, sentiment analysis, social network communities

I. INTRODUCTION

Big budget movies cost millions of rupees to the producers and the movies need to perform well within the first week of the release as there are back to back movie releases every Friday in the Bollywood. Movie success depends on a lot of factors but they sum up as acceptance of the movie in the masses. These days, movie reviews both pre and post, are posted on social networks. Users actively take part in the criticism process and their opinions are visible to their connections who may read them and show trust in them while deciding to watch the movie in the theater. Movie production is a costly affair, therefore there is a need to study the correlation between the social media sentiment and the success and failure of Bollywood movies viz. the Box-office collection.

In this new environmental setup, brands choose social network websites to extend their reach to prospective customers and keep engaged their existing customers by employing different promotional strategies through social media. These promotional strategies are the lifelines of the entertainment sector (e.g., movies) because of the uniqueness of their products requiring value creation by the promoters and the general public.Timely prediction of the performance of an upcoming movie can help the promoters of the movie to develop and modify the promotional strategy. They can sense the general acceptance of the movie being released by the prospective viewers. To increase the awareness of the movie and viewership, these days social media is of great help. The problem is to assess the impact of social media marketing and promotional strategy in predicting the Box-office success of the movie(s).

The model proposed in the research aims at predicting the success of Bollywood movies on the basis of their two weeks (one week before the movie release and opening week) social network sentiment score of the largest communities detected in the tweet dataset of movie Twitter handles. Friend circle community tweets act as a recommendation to the users in case of movies.

II. RELATED WORK

Social media is evolving as a new form of online media, which has the characteristics of participation of users, transparency regarding feedback, two-way conversation, quick and effective community formation and connectedness with like-minded users [1]. Social media is different from traditional media which is based on one-way broadcast communication and has little or no feedback mechanism. Social media has the ability to influence public opinion in policy making such as the case of stem cell research by enlarging the target audience beyond demographic boundaries.

Social media has an abundance of data and its predictive capabilities can be used to predict customer's taste [2]. Social media data has been studied for predicting the outcomes of conflicts, elections, public policies, stock prices, book sales etc. Social media represents the collective wisdom of the crowds and thus the predictive capability of the corpus depends on its size. Sentiments represent the user opinion about an entity and when they are posted on ubiquitous platforms like social media, their impact is on a large scale. Software technologies like data mining and Natural Language Processing are required to tap the full potential of social media sentiment analysis task.

Factors affecting the acceptance of recommendations of the social media followers were studied and the studies concluded that adaptive approaches give better results [3], [4]. Social media can be viewed as a new marketing mechanism which affects the customers. It creates the much needed timely hype amongst the customers about the upcoming movies by initiating and supporting the conversation about the movie.

Word of mouth (WOM) is defined as casual communication among customers (existing or prospective) about services and products[5]. It ranges from casual talks between two persons for promoting a particular brand focusing on an entire population domain. Advancement in networking and Internet technologies has helped Online Word of Mouth OWOM creep into electronic media communication [6]. Social network websites, discussion forums, blogs andnewsgroups are live examples of online word of mouth. Web users share their views and opinions related to events, movies, music albums, news, and stock prices with other users who are completely unknown to them. Users express positive, negative or neutral views or support somebody else's views by liking or sharing it on their online profile.

Movie rating public datasets like Netflix and MovieLens have become obsolete with the advent of Tweet based sentiment analysis which is more user-centered [7]. The old public datasets lacked attributes which are present in new movies. The authors collected more than 60,000 tweets for the movies released in the study period and conducted user-centric sentiment analysis of the dataset to accurately predict the performance of movies. Another major advantage of the study over existing public databases is that it includes users who have expressed their opinions about a single movie whereas the public datasets have criteria of a minimum movie rating by a user in order to be included in their dataset.

Internet Movie Database (IMDB) the movie website hosts huge datasets related to the movie domain [8]. 4 major classification algorithms on a dataset of 50,000 movie reviews from the IMDB database have been compared in this study. 25,000 reviews were used for training the algorithms and rest 25,000 were used to test them. The accuracy of the four algorithms was Naïve Bayes (86.53), Support Vector Machine (89.33), Ensemble (95.0) and Decision Tree (68.44). The study discussed the monetary benefits of movie review sentiment analysis and the final outcome of the study was that a selective combination of algorithms yields better results compared to individual algorithms.

The first week of movie release is a major indicator of a movie's success. First-week Box-office collection of movie predicted by using clustering of features extracted from Tweets lowered the errors [9]. The authors emphasize the financial aspect of prediction systems using Twitter datasets. They developed an algorithm to predict the opening weekend Box-office collection of movies using Twitter sentiment analysis. The model is a mix of unsupervised and supervised learning and predicted daily collection of movies in their first week of release. The experiment used sci-kit learn library for Python to develop a prediction model for predicting movie revenues for the opening weekend of the movie release. Their single model prediction approach achieved an accuracy of 65% and multiple-model prediction approach got 69% accuracy.

The ratio of positive sentiment tweets to negative sentiment tweets can be used as a threshold in predicting movie success [10]. The model worked on three simple rules, if the ratio of positive to negative tweets is greater than or equal to 5, the movie is labeled as a successful movie, if this ratio is between 1.5 and 5, movie is labeled as an average movie otherwise the movie is labeled as a flop one if the ratio is less than 1.5.

Fuzzy inference system algorithm for data mining can be used for the sentiment analysis of movie Tweet datasets to classify the performance of a movie as a flop, average or a hit [11]. The method used two inputs, actor rating and different sentiment polarity scores (negative, positive or neutral). The prediction model using the hype created just before the movie release gave fairly accurate results when compared with the actual results of Box-office collections. The authors stated that data about ticket price and the number of screens could have been included (this data was not available for the movies studied in the experiment) in the research to generate more accurate results.

III. DATASET CREATION

Most of the Fortune 500 companies rely on Twitter for their Internet-based marketing campaigns [12]. As of 2018, there are approximately 326 million active Twitter users who post around 500 million tweets per day on topics covering movies, terrorism, elections, medicine, science, sports, education, business and other aspects of human life. Researchers have mined enormous tweet datasets and successfully predicted the outcomes of elections, public policies, fiscal policies, civil wars and marketing trends.

Python was developed in the latter half of the 1980s by Guido van Rossum in the Netherlands as a replacement of the ABC language [13]. It has more than one paradigms language supporting object-oriented programming scheme, structured programming, functional programming andaspect-oriented programming. Since January 2017, Python is being hosted on GitHub. It is open source language and available for both Windows and Linux platforms and is supported by many Integrated Development Environments (IDE) like Jupyter, PyCharm. Python Package Index (PyPI) supports over 130,000 packages for third-party software covering multimedia, databases, graphical user interfaces, web interfaces, image processing, data analytics and visualization.

Twitter has different APIs for various types of interactions with tweets. Search APIs are used to search tweets corresponding to the search criteria which can be a hashtag (#) or specific keywords. REST APIs are used to get access to user objects, timelines and status. They are used to perform GET and POST data functionalities on Twitter. For the research experiment,a dataset was created by mining Twitter for the hashtags of the Bollywood movies released between 1st January 2017 to 31stMay 2017 using Python libraries and OAuth credentials. Movie hashtag tweets were mined for 14 days (2 weeks); one week before the movie release date and one week after the release date.For some of the movies released in this duration, authentic data could not be collected about their budget or box-office collection and hence they were excluded from the experiment. Table 1 lists the Bollywood movies sorted by their release date, box-office verdict and number of distinct tweets fetched for their Twitter hashtags. Python library functions for stemming, lemmatization and other miscellaneous text pre-processing task were used to clean the text and remove uniform resource locators, emojis and Twitter specific characters.

Table 1.Movie name, box-office verdict

Movie Name	Box-office verdict	Total tweets
Haraamkhor	Flop	254
Ok Jaanu	Flop	89
Coffee with D	Flop	178
Kaabil	Hit	1763
Raees	Hit	2820
Jolly LLB 2	Hit	2278
Runningshaadi.com	Flop	208
The Ghazi Attack	Flop	67
Irada	Flop	170
Rangoon	Flop	158
Commando 2	Hit	540
Badrinath Ki Dulhania	Hit	740
Aa Gaya Hero	Flop	217
Phillauri	Hit	418
NaamShabana	Flop	392
Begum Jaan	Flop	276
Noor	Flop	250
Baahubali 2: The Conclusion	Hit	4020
MeriPyaariBindu	Flop	208
Sarkar 3	Flop	1519
Half Girlfriend	Hit	539
Hindi Medium	Hit	1201
Sachin: A Billion Dreams	Hit	1171

When Twitter users reply to other users, retweet or share the tweets, it creates online word of mouth which leads to the formation of network communities. The “number of tweets” column of Table 1 indicates the tweets forming communities in the dataset. These communities were detected using the Louvain clustering algorithm implementation in Python [14]–[16].

The number of communities in a graph also depends on the total number of vertices in that graph[17]. In this case, each vertex is a Twitter ID whose associated user tweeted using the Twitter handle of the movie. A lesser number of vertices lead to lesser number of communities and hence a sparse network is generated. Movies which failed to capture the attention of the general public got a lesser number of tweets and the spread of information in the initial stage was slower and restricted by the number of users who initiated the opinion sharing process.

IV. SENTIMENT ANALYSIS

Sentiment analysis aims to capture opinions, views, sentiments contained in the text and categorize the source text based on the polarity [18], [19].Social network sentiment analysis is finding crucial applications in recent developments[20]. Researchers have used Twitter sentiment analysis in the ongoing Syrian refugee crisis [21]. Social media offers a free platform for people to share their feelings with others. The study used 2381297 tweets in English and Turkish language. The sentiments conveyed in Turkish language tweets were found to be different from that in the English language tweets as Turkey had opened up its borders to a large number of Syrian refugees. Turkish tweets were more positive and English tweets were more neutral in sentiments towards the refugees.

Sentiment analysis on tweet dataset is mainly a process to accurately classify the tweets into various sentiment classes [13]. Tweet dataset sentiment analysis is a challenging task owing to the limited tweet size. Each tweet with a maximum of 140 characters (recently increased to 280 characters), generating statements with a relatively smaller set of features. Use of slang or non-English words requires careful preprocessing of data before the sentiment analysis. Twitter allows the use of hashtags; URLs and user references need different processing than the regular content. Users post their views in a plethora of ways, using multiple languages in the same chain of tweets and using smileys or repeated words to convey emotions.

Python NLTK and TextBlob libraries were used to perform sentiment analysis for the entire dataset. The “sentiment” function returns a tuple of the form (polarity, subjectivity) where polarity is a float value within the range [-1.0, 1.0] where -1.0 is a negative sentiment, 0 is a neutral sentiment and 1.0 is a positive sentiment; and subjectivity is a float value within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective.

In the research presented, the polarity and subjectivity values are calculated for each tweet (each row of the CSV dataset file). The negative polarity values or the polarity values less than zero represent a negative sentiment and the positive polarity values or the polarity values greater than zero represent a positive sentiment. The final polarity of a tweet/ blog is calculated as the product of polarity and subjectivity value as shown in the equation.

$$Polarity\_Final = Polarity * Subjectivity$$

Based on the final polarity value i.e. Polarity_Final, highly positive sentiment tweets or blogs were labeled as P+, slightly positive sentiment tweets or blogs were labeled as P, highly negative sentiment tweets or blogs were labeled as N+, slightly negative sentiment tweets or blogs were labeled as N. Weights were assigned to the tweets as shown in table 2.

Table 2. Sentiment polarity and weights

Polarity_Final	Range				
	-1.0 to -0.5	-0.5 to -0.1	-0.1 to 0.1	0.1 to 0.5	0.5 to -1
Polarity	N+	N	NEU	P	P+
Weight	-2	-1	0	1	2

Using the polarity_final value shown in table 2, polarity weights were assigned to the tweets for the entire movie dataset. Table 3 gives a snapshot of the polarity assignment to the tweets of the "Sweetiee Weds NRI" movie dataset.

Table 3. Sentiment analysis snapshot for "Sweetiee Weds NRI" movie

Tweet	Sentiment	Polarity
May God boost u widlove,goodluck and success for ur movie #SweetieeWedsNRI #OpeningCeremonyTomorrow	Positive	P+
Congratulations #SweetieeWedsNRI rocks. Loved it.	Positive	P+
There is no story in the movie, total waste	Negative	N-
Good small budget movie this time	Positive	P
Hi everyone, don't know how this movie will perform	Neutral	Neu
Missing acting in the movie	Negative	N-

Sentiment analysis was performed on the entire movie tweet dataset on a daily basis starting one week prior to the movie release for 2 weeks. The number of tweets in a given polarity was multiplied with the corresponding weight of the polarity to calculate the total sentiment weight of the movies. Table 4 lists the number of tweets under each category and the total polarity weight of all the movies covered in the experiment.

Table 4. Sentiment analysis of the movie dataset

Movie Name	P+ (+2)	P (+1)	N (-1)	N- (-2)	Neu (0)	Total tweets	Total weight
Haraamkhor	21	43	82	65	43	254	-127
Ok Jaanu	14	17	23	24	11	89	-26
Coffee with D	17	19	72	49	21	178	-117
Kaabil	593	455	76	61	578	1763	1443
Raees	815	1023	213	277	492	2820	1886
Jolly LLB 2	664	896	116	256	346	2278	1596
Runningshaadi.com	19	54	63	16	56	208	-3
The Ghazi Attack	7	16	21	10	13	67	-11
Irada	24	42	56	19	29	170	-4
Rangoon	16	23	33	54	32	158	-86
Commando 2	105	214	64	56	101	540	248
Badrinath Ki Dulhania	214	245	119	38	124	740	478
Aa Gaya Hero	19	16	59	46	77	217	-97
Phillauri	56	185	29	64	84	418	140
NaamShabana	47	50	113	68	114	392	-105
Begum Jaan	13	28	49	72	114	276	-139
Noor	23	27	98	48	54	250	-121
Baahubali 2: The Conclusion	1028	1982	218	147	645	4020	3526
MeriPyaariBindu	6	12	58	65	67	208	-164
Sarkar 3	257	413	365	286	198	1519	-10
Half Girlfriend	87	141	156	46	109	539	67
Hindi Medium	314	287	251	148	201	1201	368
Sachin: A Billion Dreams	214	289	301	74	293	1171	268

V. PREDICTION MODEL

The proposed prediction model uses average polarity weight of tweets in the dataset (total polarity weight/ number of tweets in the dataset) for a given movie to predict its failure of success on the box-office.

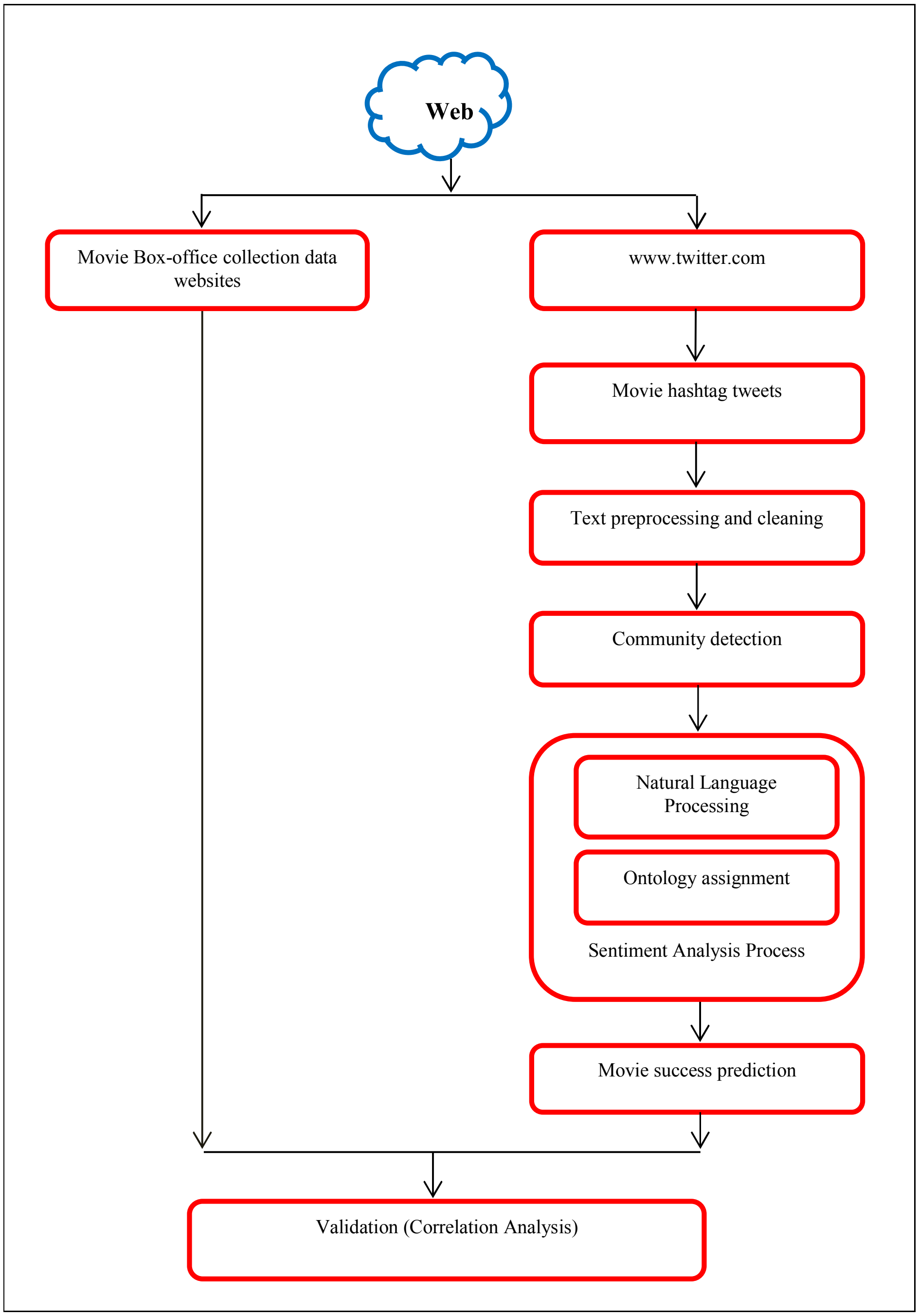

Figure 1. Community sentiment analysis based prediction model

Fig. 1 shows the architecture of the proposed mathematical model. If the average sentiment weight is less than 0, the movie is predicted to fail, if the average sentiment weight is greater than 0, the movie is predicted to succeed else if the average sentiment weight is equal to 0, the movie is predicted to be an average performer on the box-office. It is technicallydifficult to define “average movie” without fuzzy logic. Table 5 lists the verdict of the proposed model for the movies studies and their actual box-office verdict. For statistical validation, Flop has been represented as -1, average as 0 and Hit has been represented as +1 or simply 1. Out of the 23 movies, predictions for 23 movies are correct and 2 are incorrect.

Table 5. Model prediction results and actual box-office result

S. No	Movie Name	Average Tweet sentiment weight	Model prediction		Box-office verdict	
			Verdict	Numeric conversion	Verdict	Numeric conversion
1	Haraamkhor	-0.5	Flop	-1	Flop	-1
2	Ok Jaanu	-0.2921348	Flop	-1	Flop	-1
3	Coffee with D	-0.6573034	Flop	-1	Flop	-1
4	Kaabil	0.8184912	Hit	1	Hit	1
5	Raees	0.6687943	Hit	1	Hit	1
6	Jolly LLB 2	0.7006146	Hit	1	Hit	1
7	Runningshaadi.com	-0.0144231	Flop	-1	Flop	-1
8	The Ghazi Attack	-0.1641791	Flop	-1	Flop	-1
9	Irada	-0.0235294	Flop	-1	Flop	-1
10	Rangoon	-0.5443038	Flop	-1	Flop	-1
11	Commando 2	-0.0697115	Flop	-1	Hit	1
12	Badrinath Ki Dulhania	0.6459459	Hit	1	Hit	1
13	Aa Gaya Hero	-0.4470046	Flop	-1	Flop	-1
14	Phillauri	0.3349282	Hit	1	Hit	1
15	NaamShabana	-0.2678571	Flop	-1	Flop	-1
16	Begum Jaan	-0.5036232	Flop	-1	Flop	-1
17	Noor	-0.484	Flop	-1	Flop	-1
18	Baahubali 2: The Conclusion	0.8771144	Hit	1	Hit	1
19	MeriPyaariBindu	-0.7884615	Flop	-1	Flop	-1
20	Sarkar 3	-0.0065833	Flop	-1	Flop	-1
21	Half Girlfriend	0.1243043	Hit	1	Hit	1
22	Hindi Medium	0.3064113	Hit	1	Hit	1
23	Sachin: A Billion Dreams	-0.1475973	Flop	1	Hit	1

VI. VALIDATION OF RESULTS

Statistical correlation tests were used to validate the results of the study analyzing the impact of investor sentiment on stock prices [22]. User tweet sentiment was used to predict the performance of 30 company stocks forming the Dow Jones Industrial Average (Dow Jones is the New York Stock Exchange Index). For companies dealing with public or general masses, there was a high positive correlation between the social network sentiment and stock price movement. These include software companies and banks which have popularity among the general public.

IBM SPSS 20 was used to validate the predicted vs actual verdict numeric representation values of Flop and Hit performance of the movies studied in the experiment. Table 6 shows the results of Bivariate

Correlation Analysis test performed on the predicted and actual values from table 5. A high correlation of 0.833 has been achieved for the proposed mathematical model results. This validates the results of the mathematical model.

Table 6. Bivariate Correlation Analysis validation results

		Predicted verdict	Actual verdict
Predicted verdict	Pearson Correlation	1	.833**
	Sig. (2-tailed)		.000
	Number of movies	23	23
Actual verdict	Pearson Correlation	.833**	1
	Sig. (2-tailed)	.000	
	Number of movies	23	23

**. Correlation is significant at the 0.01 level (2-tailed).

VII. CONCLUSION

Social networks like Twitter contain dense communities of like-minded users spreading the online word of mouth. Dense communities originate and grow when the information being shares among the constituent nodes is of a particular interest and draws attention. This OWOM can be treated as the collective wisdom of the crowds and provides general acceptance of the masses for brands and services like movies. Sentiment analysis of Twitter mined dataset community can be used to develop mathematical models that can accurately predict success or failure of Bollywood movies. This can help the movie promoters to modify their promotion strategies to create a positive sentiment in public leading to higher revenue generation.

REFERENCES

[1] A. Adams, G. Lomax, and A. Santarini, Social media & stem cell science: Examining the discourse, Regenerative Medicine, 6(6 SUPPL), 2011, 121-124.

[2] S. Coyne, P. Madiraju, and J. Coelho, Forecasting Stock Prices Using Social Media Analysis, in 2017 IEEE 15th International Conference on Dependable, Autonomic and Secure Computing, 15th International Conference on Pervasive Intelligence and Computing, 3rd International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), 2017, 1031-1038.

[3] M. von Hoffen, M. Hagge, J. H. Betzing, and F. Chasin, Leveraging social media to gain insights into service delivery: a study on Airbnb, Information Systems and e-Business Management, 2017, 1-23

[4] Y. Liu, Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue, Journal of Marketing, 70(3), 2006, 74-89.

[5] S. Ranjan, S. Sood, and V. Verma, Twitter Sentiment Analysis of Real-time Customer Experience Feedback for Predicting Growth of Indian Telecom Companies, in 2018 4th International Conference on Computing Sciences (ICCS), 2018, 166-174.

[6] A. Giachanou and F. Crestani, Like it or not: A survey of Twitter sentiment analysis methods, ACM Computing Surveys, 49(2), 2016, 28:1-41.

[7] S. Dooms, T. De Pessemier, and L. Martens, MovieTweetings: a Movie Rating Dataset Collected From Twitter, in Workshop on Crowdsourcing and human computation for recommender systems, CrowdRec at RecSys, October, 2013, 43–44.

[8] S. K. Janane, M. S. Keerthana, and B. Subbulakshmi, Hybrid Classification for Sentiment Analysis of Movie Reviews, International Journal of Engineering Sciences and Research, 7(4), 2018, 724-728.

[9] S. Shim and M. Pourhomayoun, Predicting Movie Market Revenue Using Social Media Data, 2017 IEEE International Conference on Information Reuse and Integration (IRI), 2017, 478-484.

[10] V. Jain, Prediction of Movie Success using Sentiment Analysis of Tweets, in 2013 International Conference on Soft Computing and Software Engineering, 3(3), 2013, 308-313.

[11] D. D. Gaikar, B. Marakarkandy, and C. Dasgupta, Using Twitter data to predict the performance of Bollywood movies, Industrial Management & Data Systems, 115(9), 2015, 1604-1621.

[12] K. Swani, G. R. Milne, C. Cromer, and B. P. Brown, Fortune 500 Companies' Use of Twitter Communications: A Comparison between Product and Service Tweets, International Journal of Integrated Marketing Communications, 5(2), 2013, 47-56.

[13] B. Gupta et al., Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python, International Journal of Computer Applications, 165(9), 2017, 0975-8887.

[14] P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, Generalized Louvain method for community detection in large networks, in International Conference on Intelligent Systems Design and Applications, ISDA, 2011, 88–93.

[15] S. Dutta, S. Ghatak, R. Dey, A. K. Das, and S. Ghosh, Attribute selection for improving spam classification in online social networks: a rough set theory-based approach, Social Network Analysis and Mining, 8(1), 2018.

[16] S. Ranjan and S. Sood, Analyzing Social Media Community Sentiment Score for Prediction of Success of Bollywood Movies, International Journal of Latest Engineering and Management Research, 3(2S), 2018, 80-88.

[17] S. Ranjan and S. Sood, Identifying network communities in Bollywood Tweet dataset, International Journal of Advance Research in Science and Engineering, 6(1), 2017, 503-512.

[18] W. Medhat, A. Hassan, and H. Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal, 5(4), 2014, 1093-1113.

[19] T. Wilson, J. Wiebe, and P. Hoffman, Recognizing contextual polarity in phrase level sentiment analysis, Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, 2005, 347-354.

[20] F. L. Cruz, J. A. Troyano, F. Enríquez, F. J. Ortega, and C. G. Vallejo, 'Long autonomy or long delay?' The importance of domain in opinion mining, Expert Systems with Applications, 40(8), 2013, 3174-3184.

[21] N. Öztürk and S. Ayvaz, Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis, Telematics and Informatics, 35(1), 2018, 136-147.

[22] B. Dickinson and W. Hu, Sentiment Analysis of Investor Opinions on Twitter, Scientific Research Publishing, 4(3), 2015, 62-71.