

Towards Content-Dependent Social Media Platform Preference Analysis


Parmeet Kaur, Jaypee Institute of Information Technology, Noida, India

Shubhankar Gupta, Jaypee Institute of Information Technology, Noida, India

Shubham Dhingra, Jaypee Institute of Information Technology, Noida, India

Shreeya Sharma, Jaypee Institute of Information Technology, Noida, India

Anuja Arora, Jaypee Institute of Information Technology, Noida, India

 <https://orcid.org/0000-0001-5215-1300>

ABSTRACT

Social media is one of the major outcomes of progressive changes in the world of technology. The various social webs and mobile technologies have accelerated the rate at which information sharing is done, how relationships developed, and influences are held. Social media is increasingly being used by the people to help and shape the world's events and cultures with the ability to share pictures, ideas, events, etc. Further, it has transformed the way the authors interpret life and the way business is done. This article presents a decision system for selecting an appropriate social media platform (such as Facebook or Twitter) to post content with the objective to maximize the reachability of the post. The decision is made considering the domain or subject of the post and retrieving data associated with it from the web at regular time intervals. The retrieved data has been trained using logistics and K-NN regression to classify a particular instance of data and identify the platform which can provide the most reachability. The system also suggests keywords related to the topic of the post which has been mostly used in recent times.

KEYWORDS

Facebook, K-NN Regression, Logistics Regression, Social Media, Statistical Model, Twitter

1. INTRODUCTION

In today's modern era where the state of media is changing constantly, social media has gained immense popularity and social media revolution has come into play (Gupta and Nitin, 2017). Social media involves blogs, forums and various other aspects of interactive presence that enable the individuals to engage in conversations or discussions over a particular news article, blog post or event. As a result, social media provides a way to increase reachability of ideas, views or content of any other form. Lots of research works have been executed to solve problems in the domain of online social media (Dey, Borah, Babo, & Ashour, 2018). This includes the studies to compare hashtags used in Instagram and Twitter (Highfield and Leaver, 2015), examine the demographics of social media users (Davenport et al., 2014), determine challenges when accessing the Twitter data (Kelley et al., 2013), differentiate users on how they react on Instagram and examine the full photo content (Hu et al., 2014; Mittal et al., 2017), understand people's behavior through their speech of text on social media (Schwartz and Ungar, 2015), analyze the comments made by the user and his friends (Ko et al., 2014), to compare

DOI: 10.4018/IJACI.2020040102

Copyright © 2020, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

various tools and techniques for social media analytics (Batrinsa and Treleven, 2015) and recommend keywords to the users (Beliga et al., 2015).

A problem that has not been much discussed yet is regarding the selection of a social media platform most appropriate for a specific content to be posted. Each social media content need not be posted on all social media sites and usually, a content uploader is unable to decide the following while posting content- Which social media platform is most desirable to post different content? What content should be posted, i.e., which terms or keywords can boost the online spread of a post or increase the post engagements? Therefore, this paper proposes an approach to determine the most suitable social media platform (Facebook/ Twitter) based on post content and also the most likable content (or key terms) to ensure a wide and quick spread of a post, i.e. to make the post viral.

According to the studied literature, few research gaps exist in relevance to the objective of this paper, i.e., in the direction to enhance the influence of a post. There has been a statistical analysis in some papers related to the number of likes, comments, shares, retweet, status count and others (Davenport et al., 2014; Schwartz and Ungar, 2015; Ko et al., 2014; Kelley et al., 2013). Whereas, the other papers have reported their findings based on the surveys conducted on a sample of users. These gaps require an effective approach to identify the most relevant social media platform and keyword suggestion approach for influence maximization. A comparative analysis of various social media applications is presented in the existing literature (Ko et al., 2014). Davenport et al. worked on demographic details to find out the most preferable social media platform (Davenport et al., 2014). Therefore, based on this studied literature, we identified a few research gaps which motivated to work further in this direction. No previous work exists to identify of most preferable social media platform based on historical posts content of similar theme and users' reaction on these posts.

This research work is done to get the answer to the following research questions:

RQ1: Is it possible to measure the most preferable social media application for a specific theme/ topic? Or which social media application will make the topic more popular as compared to other social media?

RQ2: Is it possible to measure theme influence based on users' engagements and keywords used in the discussion of the post?

In previous works, keyword/ theme-based influence identification has been performed based on hashtags (Highfield and Leaver, 2015), single term frequency (Beliga et al., 2015) and n-gram word extraction, etc. However, a single approach solely will not be able to handle this issue.

RQ3: What all information retrieval and Natural language processing techniques can be used to find out the most preferable keywords for a particular post to increase its popularity index?

This paper primarily focuses on two social media platforms, namely Facebook and Twitter. This paper works on the cross-domain, i.e., compare the various parameters from the aforementioned social media platforms. The results provided for a certain query entered by the user is the statistical analysis of how the social media platforms differ in certain parameters. This will eventually help others to visualize the different scenarios in which the social media platforms differ. For example, in the case of "Bollywood", users are bound to find more information on Facebook than on Twitter. Whereas, when it comes to a particular celebrity, more information can be found on Twitter, and it's obvious since celebrities are more active on Twitter than on Facebook (Arora et al., 2019). When it comes to politics, one can find more information on Facebook. But when it comes to a specific national political party, such as BJP and Congress, one can find more information on Twitter. The papers reviewed have not been able to provide a comparative study, rather papers focus entirely on one of the social media platforms.

In context of the identified research gaps, this study makes the following two-fold research contributions to overcome identified research questions:

- Recommend the best social media platform out of Facebook and Twitter based on the historical data/ user's engagement (number of likes, shares, comments, retweets and followers). User's engagement on a particular and similar theme post on varying platforms is computed to get the most influential platform;
- Suggest the most appropriate and influential keywords which comprise of hashtags, single and n-gram keywords to the users.

The paper is organized as follow: Section 2 discusses the related and state-of-art work done by researchers in this direction, this section is further divided into three subsections according to contribution. Overall content preference analysis framework which shows the functionality of the system is depicted in Section 3. Section 4 discusses about the dataset and its collection process. Further, section 5 depicts the overall influential social media platform identification and influential keyword extraction process. Section 6 is about the validation results of proposed approach which is followed by concluding remark section.

2. RELATED WORK

The literature is studied in varying direction to explore the depth of research work done in determining the most preferable social media application selection. The studied literature is divided in three subparts to provide in-depth knowledge of work for the same. First part discusses the available tools, the second parts includes papers about the insights of approach used for most influential social media platform selection (Davenport et al., 2014, Hu et al., 2014, Schwartz and Ungar, 2014), and third subsection includes papers which focus on most influential keywords extraction of a specific theme (Beliga et al., 2015 and Bharti et al., 2017).

2.1. Existing Tools

Some Tools similar to our targeted research objective exist in market, these tools are doing partial jobs to get preferential social media platform and associated keywords. Some tools are BuzzSumo¹ (Rayson, 2016), Sociograph², Komfo³, Twitter Analytics⁴, and TweetStats⁵. The purpose of these tools is to provide the users with the results in accordance to the query made such as number of likes, comments, retweets, share of the photos and textual information posted. However, many of the tools primarily focuses on only one of the social media platforms, for instance: Sociograph provides analytics for Facebook pages and groups. Instagram Insights provides information on the followers, when they are online and more. A user can also view insights for specific posts and stories that he created to see how each performed and how people are engaging with them. Twitter analytics helps the users to richly represent their content on Twitter as well as provides analytics to measure the effectiveness. But none of the tools mentioned above works on the cross domain i.e. analyze and compare data on multiple social media platforms.

There are various tools that have been developed for analyzing the data on social media platforms- Instagram, Facebook and Twitter. Indeed, Facebook and Twitter are highly used social media platforms to spread a news/post. Facebook is currently the world's largest social network, with more than 2 billion monthly active users. Using the platform, users can post text, video and images. On the other hand, Twitter is a micro-blogging platform which allows individuals as well as groups to stay connected via exchange of short status messages, with a character limit of 280. Therefore, the literature study and research work both move around these two social media sites.

2.2. Influential Social Media Platforms

Davenport et al. stated that Twitter is the preferred means of active usage among narcissists in the college sample, and on the other hand, the adult sample prefers Facebook. It has presented the statistical results based on the FB status, Tweets, number of followers, friends to draw a parallel between Facebook and Twitter (Davenport et al., 2014). It has been able to include a large sample size on both the platforms and diverse cross section of participants. Also, the inclusion of the weights has been able to compare the relative importance of Facebook versus Twitter. However, it has not provided reasons in detail for why the adults ample prefer the Facebook, discussing only the reasons of usage of Twitter for narcissism by the college sample. Also, there has not been a balance of radical diversity, as the college sample in the paper's investigation is populated with Caucasian participants.

Kelly and his team have discussed various aspects related to conducting research on Twitter (Kelley et al., 2013). It mainly discusses the various challenges posed to the researchers when accessing the Twitter data. The various limitations relating to the number of tweets accessible by the Twitter API, user sampling and filtering, as well as legal and ethical concerns have been discussed. In addition to that, the paper also discusses the accessibility of the full tweets, log data, types of users and the private data. It also proposes some guidelines pertaining to the research of the Twitter data which can be adopted by the communities in the near future. It, however, has been not able to provide any sort of statistical analysis that could have been conducted on the accessibility of the tweets (number of tweets) accessed by the API, number of API calls permissible in a particular interval of time and amount of user information accessible. If this analysis would have been conducted, then a clearer picture could have been present as of why there is a need for guidelines for Twitter data.

The work done by Hu et al. in 2014 presents a complete study on people reaction on Instagram and examines the full photo content (Hu, Manikonda, & Kambhampati, 2014). It has also studied how the photo content differentiates between users. The authors wanted a clear vision of what types of photo are being posted on Instagram and how does it differentiate users with one another. The results show that there are 8 popular photo categories, 5 kinds of users based on the photo content posted and the audience / followers do not get influenced by the type of photo a user post.

Taneja et. al. in 2019 discussed that tradition recommender systems are not suitable to handle the issue raised due to social media applications. These social media applications suggest/ recommend products to a user based on their past preferences and these user preferences in latest recommender systems are considered as contextual dimensions (Taneja & Arora, 2019). Human sentiment is another aspect which also needs to be studied while finding out the most preferable social media application to post content. But human sentiment classification in a massive amount of social media data is a cumbersome task and tough due to the traditional NLP approaches. Ali et al. has applied Self-organizing map model, Principal Component analysis and deep learning to get promising Human sentiment classification outcome in social media data (Ali et al., 2019).

Schwartz and Ungar, (2014) wanted to explore people's behaviour through their speech of text. For this, he had chosen two most popular social media platforms i.e. Facebook and Twitter to get communication people to do by updating status, posting about their feeling and behaviour and tweeting. The paper also talked about different dictionaries such as Manual, Crowd-sourced and Deriving dictionary to get insights from the token/text. The paper also talked about different statistical techniques used to give weights to different words according to their importance. The idea of viewing how people think through their text on social media is interesting but consideration only text is a very bad idea. Analysis of images and audio files is, of course, much harder, but offers potentially complementary insights into people's thoughts and concerns (Matallah, Belalem, & Bouamrane, 2017).

2.3. Influential Content

Beliga et al. surveyed methods and approaches for keyword extraction (Beliga, Mestrovic, & Martincic-Ipsic, 2015). The systematic review of methods was collected which eventually resulted in a comprehensive review of existing approaches (Wang & Wang, 2019). Work related to keyword

extraction was explicitly mentioned for supervised and unsupervised methods, with a special emphasis on graph-based methods. Various graph-based methods were analyzed and compared.

In 2017, Bharti et al. proposed a hybrid approach to extract keyword automatically for multi-document text summarization in e-newspaper articles. The performance of the proposed approach was compared with three additional keyword extraction techniques namely, term frequency inverse document frequency (TF-IDF), term frequency adaptive inverse document frequency (TF-IDF), and a number of false alarm (NFA) for automatic keyword extraction and summarization in e-newspapers articles for better analysis (Bharti, Babu, and Pradhan, 2017).

Lai and To presented a novel grounded approach for analyzing, extracting and understanding social media content. The biggest hindrance to the use of social media is the lack of versatile technologies for the selection, processing and analyzing of the content being posted on the social media. In order to bridge the gap between the availability of user-generated raw data and the contextual information of aggregated data, the study conducted in the paper gave a grounded theory approach to analyze social media content to identify the underlying factor structure of the gathered information (Lai and To, 2015). The proposed methodology has been divided into four phases, namely, definition of goal and scope, data collection, data transformation, and results interpretation and can be used in decision support applications, such as crowd sourcing, profiling, web mining, social reputation modeling and social recommendations etc. which was indeed a great attempt. The idea was quite different in its approach and can be considered to be fruitful for electronic commerce researchers, organizations, and governments to understand the commonality in the online text data that appears in social media. Through the obtained information, researchers can understand the beliefs, attitudes, and perceptions of social media users with respect to the usage of user-generated content.

A summarization of work done in this research direction to predict the most preferable social media application and to find out keyword to make content popular is shown in Table 1.

3. CONTENT PREFERENCE ANALYSIS FRAMEWORK

This research work presents a comparative overview of social media platforms (Facebook and Twitter) concerning post influence. This shall eventually help the users to decide which platform is ideal for the domain of the post made to maximize post's influence. Thus, users will be able to maximize a post's reachability not only amongst their social contacts but also amongst the masses. This work has been performed in multiple steps as shown in Figure 1.

Initially, in step1, content is retrieved from two social media platforms, i.e., Facebook and Twitter. Therefore, Twitter and Facebook API are used to retrieve content in the JSON format for a particular query/term. The Facebook post data contains- number of posts, likes and shares count in accordance with the theme entered by the user. Similarly, the tweet content from Twitter consists of a statistical count of tweets, followers, and retweets. The second step is to find out the influential platform for a specific theme post. For this, a statistical model is proposed which computes influence on social media sites over a varying period for a specific post. It is observed that influence varies according to time. This objective is achieved by the application of Logistics and K-NN regression on the content retrieved at different time instances to predict the outcome.

The third and last step is to retrieve influential keywords for a specific theme post. This might help netizens to decide what keywords have more weight age as compared to others to boost a post. Various information retrieval approaches – hashtags extraction, N-gram, connotations, etc. are fetched from the content of the influential platform along with their frequency count in the decreasing order. The overall framework is depicted in pictorial form in Figure 1.

Table 1. Summarization of relevant literature

Author and Reference	Technique Used	Dataset	Results	Limitations
Davenport et. al., 2014	Statistical Techniques to analyse preferable platform and draw a user connection between Twitter and Facebook	Focused group: College students and Adults students	Twitter preferred by college students and Facebook preferred by adults	- Not having balance of radical diversity - Machine learning results not validated
Chandok & Kumaraguru, 2017	User identification across social networks applied weighted sum method and probabilistic method	User profile attributes along with identity pair 23,985 identity pairs of same users on Twitter and Facebook	- Linkability Score	- Due to API's restrictions, only a limited number of profile features are used
Zafarani, & Liu, 2013	cross-media user identification problem. Introduced MOBIUS methodology which uses various classifiers such as Decision Tree, Random forest, Logistic Regression, SVM	websites where users have the opportunity of listing their identities (user accounts) on different sites.	93% accuracy in user identification	- Discovering features indigenous to specific sites
Kumar, Zafarani, & Liu, 2011	Working on site migration, attention migration, user activities to study user migration patterns	17,798 user profiles on 7 social media sites	4575	5663917

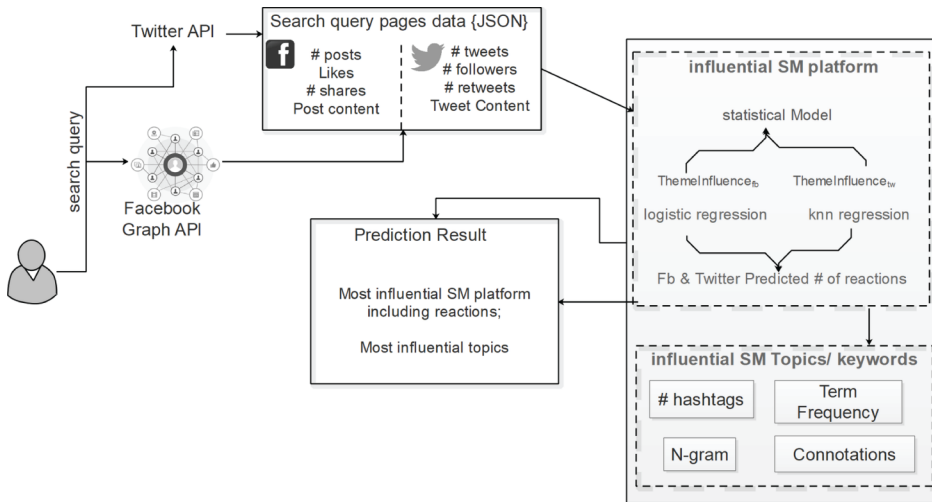
4. DATASET COLLECTION

In recent years, APIs provided by different social media platforms have been beneficial for the researchers to access and analyze data with ease. Likewise, the data mentioned in the paper has also been extracted by using APIs i.e. Facebook Graph API and Twitter API provided by Facebook and Twitter respectively. The Graph API provides an access token, which is unique to the user to extract data in the JSON format. For a particular topic entered by the user, say IPL, the Graph API returns the *page_id* of the pages pertaining to the topic in sequential order. From all the *page_id*, a page is chosen which is verified by Facebook (with a blue tick next to the title). For example, for the above-mentioned topic, the URL of the selected page is <https://www.facebook.com/IPL>. Three features which have been extracted for quantitative analysis were:

1. **Number of posts:** Posts made by the page owner and other users;
2. **Likes:** The number of users who have liked the page;
3. **Number of shares:** Total number of times page is shared by users.

Similarly, the data of the most relevant page on Twitter following the topic (<https://twitter.com/IPL>) is retrieved in JSON format using a token, called as OAuth token. For Twitter, the three features extracted for quantitative analysis were:

Figure 1. Content preference analysis framework



- **Tweets:** The count of tweets made by the page owner and other users;
- **Followers:** The number of users following the page;
- **Retweets:** The count of sharing tweets by the users.

Apart from these features, all the posts made on the Facebook Page and all the tweets made on the Twitter page has also been extracted to analyze the text for keyword recommendation i.e. which keywords have been used often along with their frequencies. The paper has only focused on three features so as to draw a parallel between the two leading social media platforms. Since the data is large and is in JSON format, it has been stored in MongoDB warehouse. The proposed approach outcome is validated on two broad theme posts on Facebook and Twitter. The Data statistics of IPL and BJP are presented in Table 2 and Table 3 respectively for both Facebook and Twitter social media applications.

5. INFLUENTIAL SOCIAL MEDIA PLATFORM IDENTIFICATION

5.1. Statistical Model

With an increase in the number of social media platforms where each platform comes with its own features and limitations, the idea of What, Where and How to express the thoughts in a perfect manner

Table 2. Data statistics for IPL

	FB Followers	FB Shares	FB Number of Posts	Twitter Followers	Twitter Shares	Twitter Number of Posts
TP 1	20869746	417734	4550	5658750	90	51575
TP 2	20869807	412077	4550	5658925	93	51576
TP 3	20874239	401315	4575	5663847	98	51636
TP 4	20874287	401315	4575	5663917	101	51637

Table 3. Data statistics for BJP

	FB Followers	FB Shares	FB Number of Posts	Twitter Followers	Twitter Shares	Twitter Number of Posts
TP 1	14308844	465289	4775	9645614	5	169336
TP 2	14310282	546042	4775	9645614	5	169398
TP 3	14310355	546042	4777	9645614	5	169399
TP 4	14310508	546044	4780	9645614	7	169409

is yet more confusing. Hence, based on the data collected using the APIs of the respective social media platforms, a statistical model has been built.

Theme influence on Facebook for a particular theme is computed using Equation 1 and Equation 2:

$$ThemeInfluence_{fb} = \sum_{i=1}^n weightPage_i \quad (1)$$

where:

$$weightPage = W_1 * Post_{count} + W_2 * likes_{counts} + W_3 * Shares_{count} \quad (2)$$

and $w1=0.7$, $W2=0.1$ and $W3=0.2$.

Theme influence on twitter for a particular theme is computed using Equation 3 and Equation 4:

$$ThemeInfluence_{tw} = \sum_{i=1}^n weightPage_i \quad (3)$$

where:

$$weightPage = W_1 * Tweet_{count} + W_2 * Followers_{count} + W_3 * retweet_{count} \quad (4)$$

and $w1=0.7$, $W2=0.1$ and $W3=0.2$.

5.2. Regression Models and KNN-Regression

Post/ tweet influence based on users' engagement is classified by two well-known machine learning classification algorithms- Logistic Regression and Knn-Regression (Dey, Wagh, Mahalle, & Pathan, 2019). These classification algorithms input user's engagements and classify the most influential social media platform for a specific theme post based on users' engagement on historical similar theme content.

Statistical model results (solution approach 1) are compared with logistic and KNN-regression results. Hence, both algorithms are used to distinguish the role of user's engagements over Facebook's post or twitter's tweets and also validate the solution approach 1 results.

5.2.1. Logistic Regression

Logistics Regression is a widely used classification algorithm which finds the probability of Success/Failure of a particular event. For a given set of independent input user engagements, logistic regression outcomes success/ failure, where success stands for ‘Facebook’ is influential and failure stands for ‘Twitter’ is influential. Basically, it predicts the probability of an influential social media portal by fitting data to a logit function. Logit function’s variable represents a probability p and logit function gives the log-odds i.e. logarithm of odds as shown in Equation 5, where odds is probability of event occurrence/ probability of event denial:

$$odds = \frac{p}{1 - p} \quad (5)$$

Logit function of number p lies in between 0 and 1 where $p=1$ means Facebook and $p=0$ means twitter, logit function formula is given in Equation 6:

$$\text{logit}(p) = \log\left(\frac{p}{1 - p}\right) = b_0 + b_1x_1 + \dots + b_kx_k \quad (6)$$

where, x_1, x_2, \dots, x_k are independent variable representing posts, likes and share count for Facebook and tweets, followers and retweet for twitter. For both, k is the number of independent variables and value of k is 3.

5.2.2. KNN-Regression

KNN-regression is another widely used classification algorithm and it is a non-parametric regression i.e. no regression function exists in this. KNN-regression classifies an object based on the majority votes of its k most similar instances. To find out the nearest neighborhood according to user engagements, distance metric between two data points needs to be computed. A popular choice Euclidean distance has been used to find out nearest neighbor as shown in Equation 7:

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad (7)$$

Hence, KNN-regression consists of a prediction point X_0 and k training observation closest to X_0 . Accuracy of KNN-regression depends on the value of k . For obtaining an optimal k value, cross validation score is calculated using $k\text{-fold} = 5$ to minimize misclassification error rate. KNN has higher accurate prediction power as compared to linear regression because it takes care of data non-linearity values as well.

5.3. Influential Social Media Topic / Keyword Identification

The choice of words on part of a user decides what words might have a huge impact as well as create an everlasting memory on others or not. Hence, influential keywords are extracted based on the historical content from the influential and maximally engaged tweet/post. Before any analysis, it is necessary to preprocess the unstructured data and get rid of noise as social media contains high-end noisy content. This will further ensure a smooth implementation of the topic extraction technique. All the contents are passed through a text cleaning pipeline with different steps:

- **Text Encoding Standardization:** Conversion of text encoding to standard utf-8 encoding;
- **Boundary Punctuation Removal:** Removal of boundary punctuations from descriptions such as cherrs!!!!, Wow:, #csk);
- **Stop words Removal:** Removal of common terms such as - a, an, the, he, am, is, of, are, etc.;
- **Extra White Spaces and Proper Case:** Removal of extra white space in the text and converting them to lowercase;
- **Digits, Short words, and alphanumeric:** Removal;
- **Words Normalization using Lemmatization:** Conversion of all the terms into their base form using word lemmatization algorithms.

Further, it goes through the process of tokenization and stemming using the function of word_tokenize() from nltk package. Thereafter the tokens are displayed to the users along with their frequencies in the decreasing order. For the remaining terms, the part of speech distribution is applied. According to observation, we notice that NN (Noun, singular), NNS (Noun, plural), VBG (Verb, principle), VBN (verb, past) and JJ (Adjectives) are highly important as compared to any other pos tag. Also, these are the tags which define the majority of the corpus. Only important POS tags are selected and all the terms with other tags are removed.

N-gram has been used to find out the terms/ keywords contain the contiguous sequence of more than one term from a given sequence of content. We consider N-gram as a feature to match influential terms and consider N-gram of different sequence length which will start matching from largest term set (here $N=4$) sequence length till single term. For example, the query “Shane Watson” is a 2-gram term set and “Chennai Super Kings” is a 3-gram term set. We will finalize the N-gram term set as a single term on the basis of its frequency in the used taken dataset. N-Gram method has been applied to suggest appropriate and relevant keywords where ‘n’ is an integral value greater than or equal to 1. In the field of linguistics, n-gram is a contiguous sequence of n items from a given sample of textual data. The items referred here might be the hashtags and keywords.

Therefore, we consider terms successfully coming under n-gram category as a single term while maintaining the frequency of terms set at the time of identification of the most influential term/ keyword according to term similarity matching.

6. RESULTS

Though social media provides a platform for users to showcase their ideas and gain recognition, yet it is difficult to select one effective social media platform corresponding to the content to be posted out of multiple such platforms. The results of the current work aid a user to determine the platform ideal for a particular situation or a specific objective. This selection is envisaged to improve the profitability for business enterprises or individuals posting online content.

6.1. Data Visualization

As a first step, the collected data is visualized to understand the relationship between data and social media platforms. Figure 2 is a plot showing the comparison between the number of Facebook and Twitter posts related to IPL at various time intervals on a particular day. In a similar manner, Figures 3 and 4 are plots showing the comparison between the number of shares and followers of both the social media platforms respectively over a period of time on a particular day.

6.2. Influential Platform Identification Results

On the basis of the above results, the corresponding scores are evaluated and plotted for the given topic i.e. IPL and BJP, as seen in Figures 5 and 6 respectively. These scores are then used to recommend the most influential social media platform. In this case, Facebook is evidently the most influential platform.

6.3. Performance Evaluation

In KNN - Regression the best k value is found using one of the most popular validation techniques known as k -fold cross validation. In k -fold cross validation, the training set is randomly divided into k groups, or folds, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining $k-1$ folds. The misclassification rate is then computed on the observations in the held-out fold. This procedure is repeated k times; each time, a different group of observations is treated as a validation set. This process results in k estimates of the test error which are then averaged out. We set $cv = 5$ for performing 5 folds on our training set and accuracy as our

Figure 2. Comparison of number of posts on Facebook and Twitter IPL page

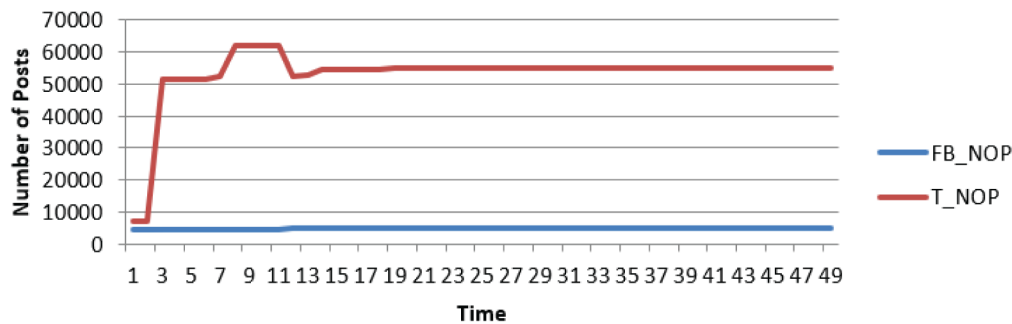


Figure 3. Comparison of number of shares of Facebook and Twitter IPL page

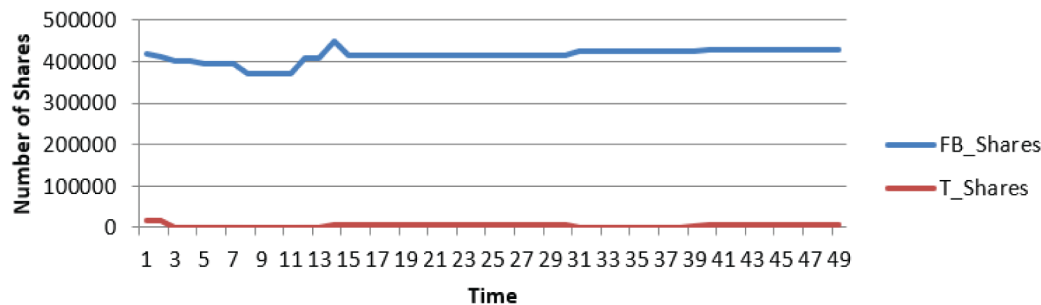


Figure 4. Comparison of number of followers of Facebook and Twitter IPL page

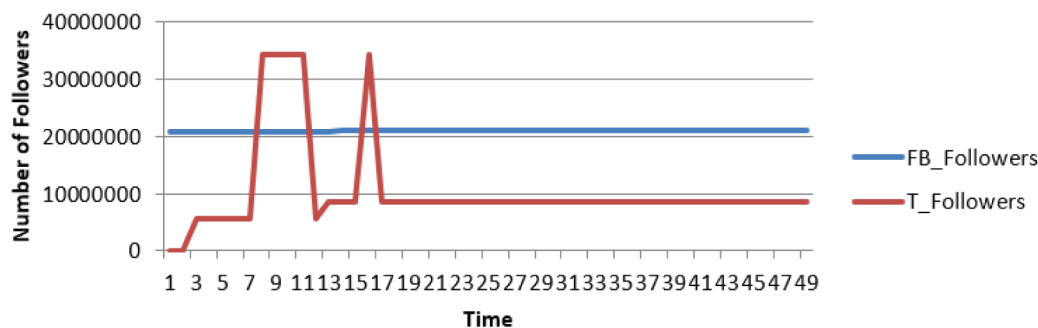
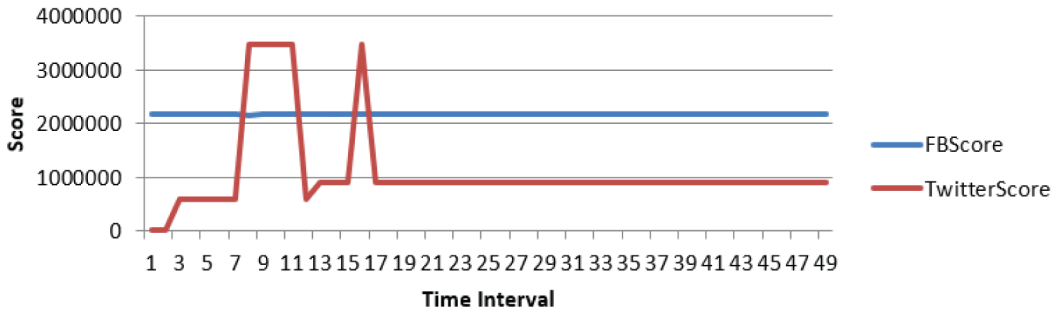


Figure 5. Statistical model theme influence results for IPL-2018



scoring metric. Finally, we plot the misclassification error versus K . As it can be seen from the obtained graph (Figure 7), the obtained k value was $k=5$ in our case.

The analysis shows that 75% of results obtained by the proposed model were correct according to LR. Similarly, 82% of the results of the proposed model were validated by analysis through KNN-regression. The precision, recall, and f-score of LR and KNN results are shown in Figure 8 (a) and 8 (b) respectively.

6.4. Influential Keywords Identification Results

The influential keyword identification results are validated for various themes and topics. Data analytics gives results of the most influential keyword of a specific theme which make it popular in the market. Top 5-5 keywords for IPL and BJP themes are depicted in Table 4. The Tool which we have developed to validate the research objective is shown in Figure 9 which shows the Facebook as a most preferable platform to post IPL related posts and present the influential keywords which user should use while posting the content on social media site.

Basically, Table 4 shows the various frequently used keywords for IPL and BJP that might be used by the user to popularize his/her post. Further, Topics those makes post influential on Facebook and twitter is shown in Figure 10 which illustrates the recommended keywords at different intervals on Facebook and Twitter for IPL theme.

Figure 6. Statistical model theme influence results for BJP

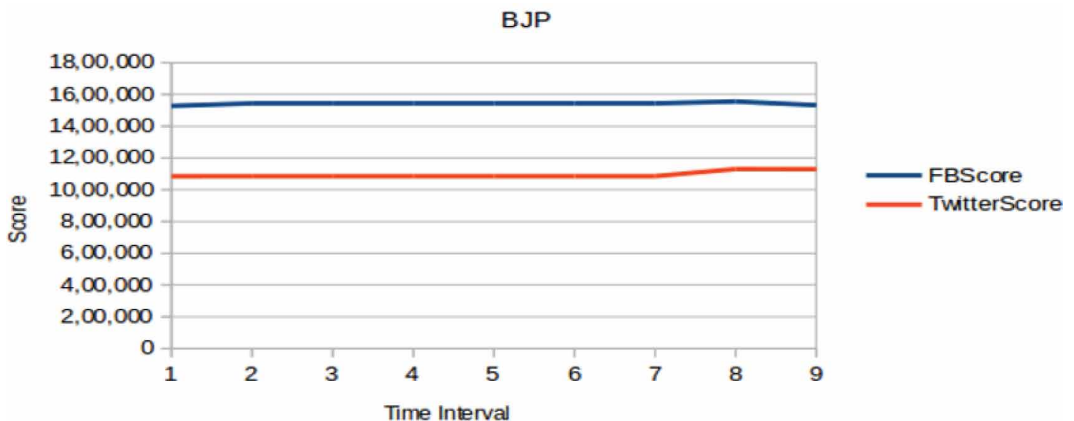


Figure 7. K-Fold KNN-Regression misclassification error graph

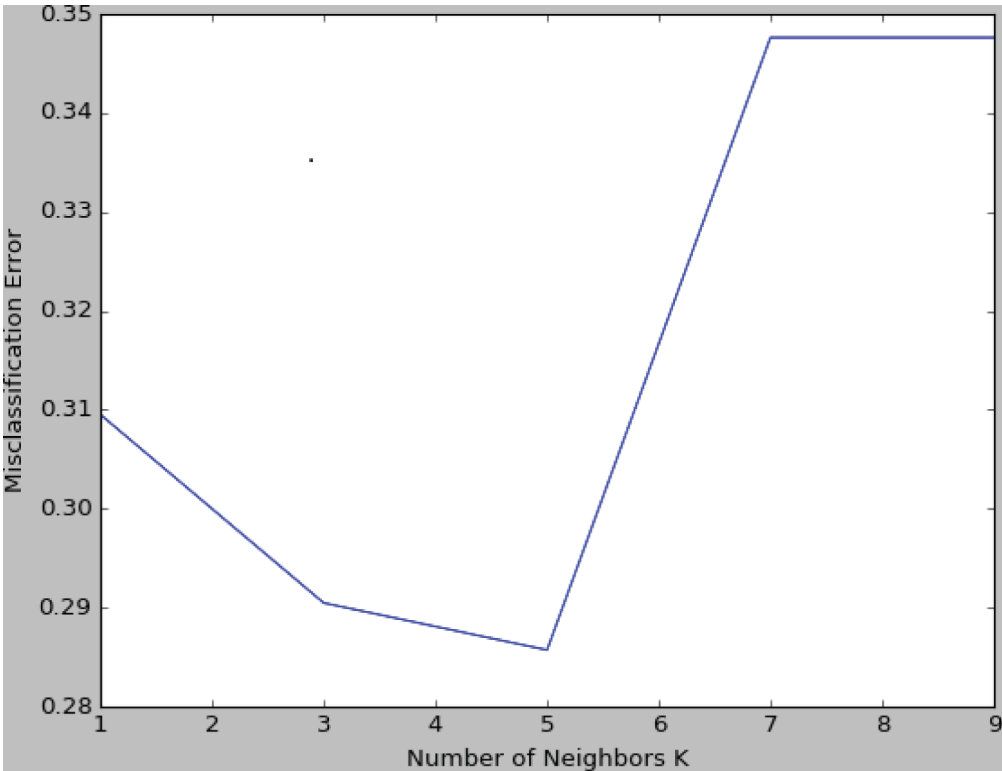


Figure 8. Logistic regression and KNN regression results

Accuracy of logistic regression classifier on test set: 0.75

	precision	recall	f1-score	support
0	0.50	1.00	0.67	1
1	1.00	0.67	0.80	3
avg / total	0.88	0.75	0.77	4

(a) Logistic Regression Result

	precision	recall	f1-score	support
0	1.00	0.25	0.40	4
1	0.81	1.00	0.90	13
avg / total	0.86	0.82	0.78	17

0.823529411765

(b) KNN Regression Result

6.5. Threats to Validity

The process to extract similar theme posts on different social media applications poses a threat to the brand (post brand) privacy. Companies may or may not desire to leak their keyword posting trend scenario and identities. Nowadays, content is viral on social media primarily due to paid post instead of organic post. Our system may mislead the user (who is going to post) as it will recommend social media platform and keywords based on viral post content which could be misleading post. Our system is not able to classify in organic and paid post which could enhance its performance in the future. The proposed system is recommending social media application and keywords based on just one content type which is text, no work has been done to ensure post virality for an image post and video post.

Figure 9. Snapshot of result for IPL query

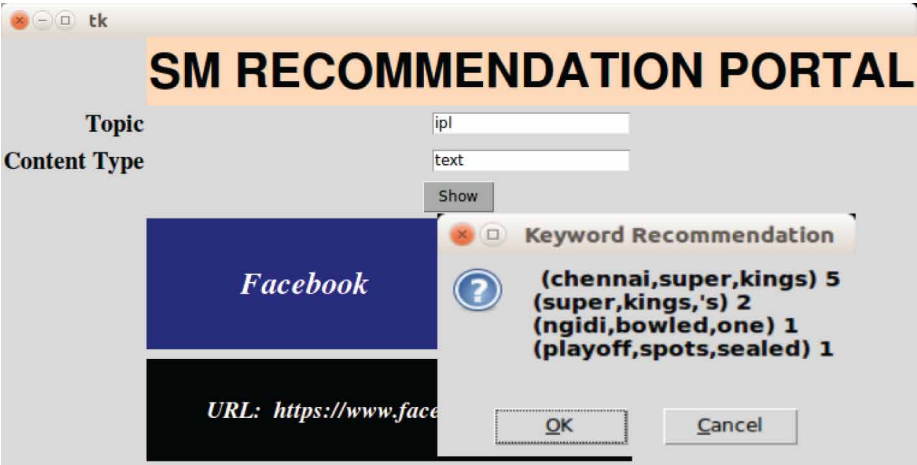
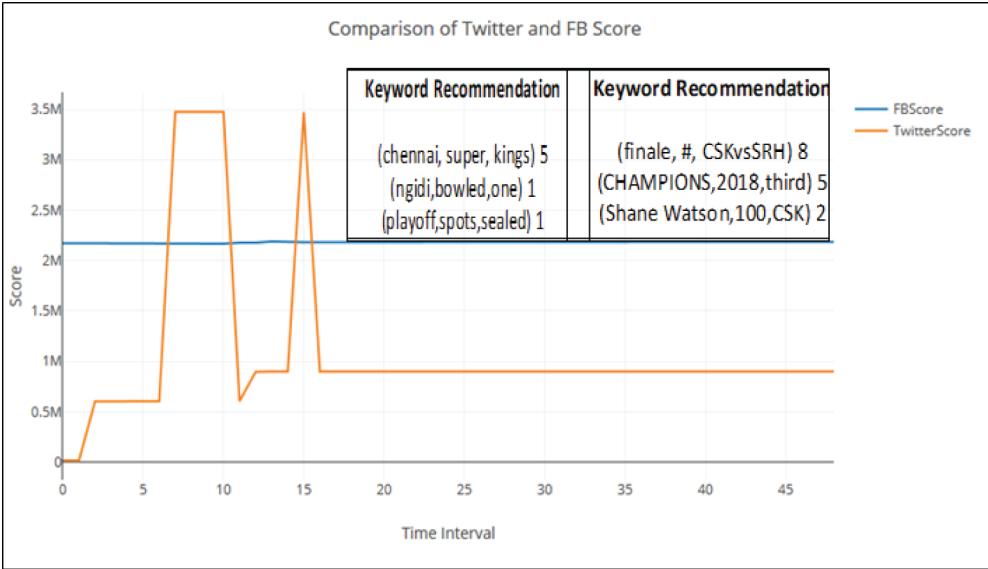


Table 4. Top 5 Most influential topics/ keywords for IPL and BJP themes

IPL	BJP
Chennai	Karnataka
Shane	Elections
Watson	JoinBJP
Playoff	bjp.org
CSK	modiat4

Figure 10. Influential keyword at different timestamp for Facebook and Twitter



7. CONCLUSION

The paper proposed an approach by means of which a user can find the most appropriate social media platform out of Facebook or Twitter in accordance with the topic of the post. Further, in order to make the post more readable, viral and popular amongst masses, the methodology of keyword recommendation has also been provided. The work considered data related to two topics, i.e., BJP and IPL. For this data, the proposed model determines the values of the features aforementioned at different time instances. Subsequently, using logistics and k-nn regression, the data has been trained successfully and also classified successfully with a high accuracy of 0.75 (for logistics regression) and 0.82 (for k-nn regression). Hence, the results demonstrate that the proposed model and keyword recommendation has performed well. The model can be used by advertising agencies for marketing products or by educational institutions and corporate firms for their advertisement or even by an individual for gaining popularity.

REFERENCES

- Ali, M. N. Y., Sarowar, M. G., Rahman, M. L., Chaki, J., Dey, N., & Tavares, J. M. R. (2019). Adam Deep Learning With SOM for Human Sentiment Classification. *International Journal of Ambient Computing and Intelligence*, 10(3), 92–116. doi:10.4018/IJACI.2019070106
- Arora, A., Bansal, S., Kandpal, C., Aswani, R., & Dwivedi, Y. (2019). Measuring social media influencer index-insights from Facebook, Twitter and Instagram. *Journal of Retailing and Consumer Services*, 49, 86–101. doi:10.1016/j.jretconser.2019.03.012
- Batrinca, B., & Treleaven, P. C. (2015). Social media analytics: A survey of techniques, tools and platforms. *AI & Society*, 30(1), 89–116. doi:10.1007/s00146-014-0549-4
- Beliga, S., Mestrovic, A., & Martincic-Ipsic, S. (2015). An overview of graph-based keyword extraction methods and approaches. *Journal of Information and Organizational Sciences*, 39(1), 1–20.
- Bharti, S. K., Babu, K. S., & Pradhan, A. (2017). Automatic Keyword Extraction for Text Summarization in Multi-document e-Newspapers Articles. *European Journal of Advances in Engineering and Technology*, 4(6), 410–427.
- Chandok, S., & Kumaraguru, P. (2017). User identities across social networks: quantifying linkability and nudging users to control linkability [Doctoral dissertation].
- Choi, D., & Kim, J., & Lee, E. (2014) Research for the Pattern Analysis of Individual Interest Using SNS Data: Focusing on Facebook. In *Proceedings of the 8th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*. Academic Press. doi:10.1109/IMIS.2014.94
- Davenport, S. W., Bergman, S. M., Bergman, J. Z., & Fearnington, M. E. (2014). Twitter versus Facebook: Exploring the role of narcissism in the motives and usage of different social media. *Computers in Human Behavior*, 32, 212–220.
- Dey, N., Borah, S., Babo, R., & Ashour, A. S. (2018). *Social Network Analytics: Computational Research Methods and Techniques*. Academic Press.
- Dey, N., Wagh, S., Mahalle, P. N., & Pathan, M. S. (Eds.). (2019). *Applied Machine Learning for Smart Data Analysis*. CRC Press. doi:10.1201/9780429440953
- Gupta, S. (2017). Detection and Elimination of Censor Words on Online Social Media. *International Journal of Computer Science and Information Technologies*, 8(5), 545-547.
- Gupta, S., & Nitin, . (2017). Development of Security Detection Model for the Security of Social Blogs and Chatting from Hostile Users. *International Journal of Computer Science & Information Technology*, 9(5), 39–49. doi:10.5121/ijcsit.2017.9504
- Highfield, T., & Leaver, T. (2015). A methodology for mapping Instagram hashtags. *First Monday*, 20(1), 1–11.
- Hu, Y., Manikonda, L., & Kambhampati, S. (2014). What we Instagram: A first analysis of instagram photo content and user types. In *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014* (pp. 595-598). The AAAI Press.
- Kelley, P. G., Cranshaw, J., & Sleeper, M. (2013). Conducting research on Twitter: A call for guidelines and metrics.
- Ko, B., Choi, D., Kim, J., Lee, E., Choi, C., Hong, J., & Kim, P. (2014, July). Research for the pattern analysis of individual interest using SNS Data: focusing on Facebook. In *Proceedings of the 2014 Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing* (pp. 36-40). IEEE.
- Kumar, S., Zafarani, R., & Liu, H. (2011, August). Understanding user migration patterns in social media. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*. AAAI Press.
- Lai, L. S. L., & To, W. M. (2015). Content analysis of social media: A grounded theory approach. *Journal of Electronic Commerce Research*, 16(2), 138–152.

Matallah, H., Belalem, G., & Bouamrane, K. (2017). Towards a new model of storage and access to data in big data and cloud computing. *International Journal of Ambient Computing and Intelligence*, 8(4), 31–44. doi:10.4018/IJACI.2017100103

Mittal, V., Kaul, A., Gupta, S. S., & Arora, A. (2017). Multivariate Features Based Instagram Post Analysis to Enrich User Experience. *Procedia Computer Science*, 122, 138–145. doi:10.1016/j.procs.2017.11.352

Rayson, S. (2017). *The Most Shared Facebook Content 2017*. The Top Viral Posts, Videos and Articles. Buzzsumo.

Schwartz, H. A., & Ungar, L. H. (2015). Data-driven content analysis of social media: A systematic overview of automated methods. *The Annals of the American Academy of Political and Social Science*, 659(1), 78–94. doi:10.1177/0002716215569197

Taneja, A., & Arora, A. (2019). Modeling user preferences using neural networks and tensor factorization model. *International Journal of Information Management*, 45, 132–148. doi:10.1016/j.ijinfomgt.2018.10.010

Wang, R., & Wang, G. (2019). Web Text Categorization Based on Statistical Merging Algorithm in Big Data Environment. *International Journal of Ambient Computing and Intelligence*, 10(3), 17–32. doi:10.4018/IJACI.2019070102

Zafarani, R., & Liu, H. (2013, August). Connecting users across social media sites: a behavioral-modeling approach. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 41–49). ACM. doi:10.1145/2487575.2487648

ENDNOTES

- ¹ <https://buzzsumo.com/>
- ² <https://sociograph.io/>
- ³ <https://komfo.com>
- ⁴ <https://analytics.twitter.com/about>
- ⁵ <http://www.tweetstats.com>

Parmeet Kaur (PhD) is currently working in Jaypee Institute of Information Technology, NOIDA as an Assistant Professor (Senior grade) and has an academic experience of over 15 years. She received her PhD (Comp Engg) from NIT Kurukshetra in 2016, M.Tech.(CSc) from Kurukshetra University in 2008 and B.E.(Hons) (CSc & Engg.) from P.E.C., Chandigarh in 1998. She is presently supervising 3 doctoral students. Her research interests include distributed systems, cloud computing, Big Data, and security.

Shubhankar Gupta completed his Bachelor of Technology in Computer Science and Engineering from JIIT, Noida in June 2018. Currently, he is pursuing master's in computer science at University of California, Davis. His areas of research interest include social media analytics, natural language processing and cyber security and has authored few research papers under the same. He has experience in developing ERP applications on Salesforce platform. He also has three months of industry experience as a Software Engineering Intern at Equinix, Inc where he developed a Full Stack Application using PostgreSQL, Express.js, React.js, and Node.js.

Shubham Dhingra was awarded his bachelor's degree in Computer Science from Jaypee Institute of Information Technology in 2018. He has co-authored a research paper titled 'Fault Tolerant Streaming of Live News using Multi-Node Cassandra' at the 2017 Tenth International Conference on Contemporary Computing (IC3), 2017 and IEEE Xplore. His main research interests include Big Data, machine learning, data mining, etc.

Shreeya Sharma is a Data Analyst in the Risk Advisory Services at Ernst & Young, Gurgaon, India. She was awarded her bachelor's degree in Computer Science from Jaypee Institute of Information Technology in 2018. She has co-authored a research paper titled 'Fault Tolerant Streaming of Live News using Multi-Node Cassandra' at the 2017 Tenth International Conference on Contemporary Computing (IC3), 2017 and IEEE Xplore. Her main research interests include Big Data, machine learning, data mining, etc. Formerly, Shreeya has coordinated and managed a team of 40 actors and theatre aspirants during her college days as a part of the theatre club.

Anuja Arora (PhD) is working as an Associate Professor in the Computer Science Engineering Department of Jaypee Institute of Information Technology. She is having academic experience of 15 years and industry experience of 1.5 years. She is Senior IEEE Member, ACM Member, SIAM Member and Life Member of IAENG. She is also a Vice-Chair for the Delhi ACM-W Chapter. She has more than 60 research papers in peer-reviewed International Journal, Book Chapter, and Conferences. Two students have been awarded Ph.D. under her supervision and three more are in process. Her research interests includes deep learning, artificial neural networks, social network analysis and mining, social media, data science, machine learning, data mining, web intelligence, web application development and web technologies, and software engineering, software testing and information retrieval systems. She is a reviewer of many reputed and peer-reviewed IEEE transactions TKDE, TNSM, IEEE transaction of cybernetics, etc. She is also the reviewer of various Springer, IGI Global, Inderscience, and De Gruyter journals. She has guided more than 17 M.tech thesis and around 100 B.tech major and minor projects.