A Predictive Methodology for Analysis of Social Media Influence in Brand Building

Pallavi Bagde
Department of Computer Science & Engineering
Medicaps University, A.B Road, Pigdambar, Rau,
Indore Madhya Pradesh 453331

Dharmendra Mangal
Department of Computer Science & Engineering
Medicaps University, A.B Road, Pigdambar, Rau,
Indore Madhya Pradesh 453331

ABSTRACT
The data mining and their techniques are classically used for analysis of data patterns. These patterns recovery is used to estimate the similar patterns on the newly obtained data. Now a day's the data mining and their techniques are frequently utilized for various business intelligence applications. The main aim of the proposed work is to recover the social media post popularity and future trends of the popularity patterns. In this context two data mining models are applied on facebook post dataset. The post dataset contains the user activities on the facebook posts published by a page manager. The user activity dataset is first processed using k-means clustering algorithm which is an unsupervised learning technique. That algorithm is applied in order to estimate which kinds of post are highly attracting users and which kind of contents are less. In addition of that for measuring the growth and future trends of a post the C4.5 (J48) decision tree algorithm is applied. By traversing the generated decision tree the post popularity trend is estimated.

The implementation of the proposed technique is performed using JAVA technology. Additionally the performance of system is measured in terms of memory and time consumption during data analysis. According to obtained results the proposed technique is effective and able to recover required data patterns from the facebook post dataset.

Keywords
Data mining, clustering, classification, trends prediction, post categorization, Dataset

1. INTRODUCTION
The exponential expanding in the quantity of web clients combined with the new improves in the media transmission advances, have made the internet based life as appropriate place for clients to examine thoughts and suppositions about administrations and items. The worldwide dissemination of social media was triggered by the exponential growth of Internet users, leading to a completely new environment for customers to exchange ideas and feedback about products and services. Branding is defined as the body processes that enhance the equity of a trade name [1] [2].

Companies soon realized the potential of using Internet-based social networks to influence customers, incorporating social media marketing communication in their strategies for leveraging business. Measuring the impact of advertisement is an important issue to be included in a global social media strategy. A system that could predict the impact of each of their advertising posts in a social media would provide a valuable advantage when deciding to communicate through social media, helping to promote products and services, thus supporting brand building and there are no studies supporting the direct relation between post form and performance [3].

Data mining provides an interesting approach for extracting predictive knowledge from raw data. Its application to social media analytics has been extensively applied, especially for evaluating future trends from users' inputs. Our study analyses and proposed a new data mining approach for classifying future trend using facebook dataset.

2. BACKGROUND
The background of a study is an important part of our research paper. It provides the context and purpose of the study. Hence there is need for background study that contribute to prepare proposed system.

2.1 What is Web Content Mining?
The Quest for knowledge has led to new discoveries and inventions. With the emergence of World Wide Web, it became a hub for all these discoveries and inventions. Today the evolution of the World Wide Web has brought us enormous and ever growing amounts of data and information [4].

Web content mining is connected however unique in relation to information mining and content mining. It is identified with information mining in light of the fact that numerous information mining systems can be connected in Web content mining. It is identified with content mining since a significant part of the web substance is writings. In any case, it is additionally very not the same as information mining since Web information are essentially semi-organized as well as unstructured, while information mining bargains principally with organized information. Web content mining is additionally not quite the same as content mining in view of the semi-structure nature of the Web, while content mining centers around unstructured writings. Web content mining accordingly requires inventive uses of information mining as well as content mining procedures and furthermore its own particular remarkable methodologies. In the previous couple of years, there was a quick development of exercises in the Web content mining zone. This isn't amazing due to the incredible development of the Web substance and critical monetary advantage of such mining [5]. Figure 1 demonstrate the web content mining information:

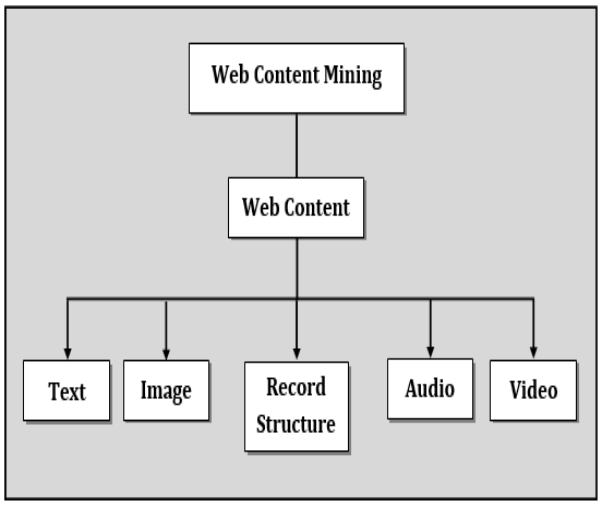

Figure 1 Web Content Data

Be that as it may, because of the heterogeneity and the absence of structure of Web information, robotized revelation of focused or unforeseen learning data still present numerous testing research issues. In this instructional exercise, we will inspect the accompanying imperative Web content mining issues and examine existing systems for taking care of these issues [6, 7].

- ✓ **Data/information extraction**: Our emphasis will be on extraction of organized information from Web pages, for example, items and indexed lists. Extricating such information enables one to give administrations. Two principle sorts of methods, machine learning and programmed extraction are secured.
- ✓ **Web information integration and schema matching**: Despite the fact that the Web contains a tremendous measure of information, each site (or even page) speaks to comparable data in an unexpected way. Step by step instructions to distinguish or coordinate semantically comparative information is an imperative issue with numerous handy applications. Some current methods and issues are inspected.
- ✓ **Opinion extraction from online sources**: There are numerous online feeling sources, e.g., client surveys of items, discussions, web journals and talk rooms. Mining suppositions (particularly buyer feelings) is of incredible significance for advertising insight and item benchmarking. We will present a couple of assignments and systems to mine such sources.
- ✓ **Knowledge synthesis**: Idea chains of command or philosophy are helpful in numerous applications. Be that as it may, creating them physically is exceptionally tedious. A couple of existing strategies that investigates the data repetition of the Web will be displayed. The primary application is to incorporate and sort out the snippets of data on the Web to give the client a sound photo of the subject space.

2.2 Social Media Analytics

Web based life Analytics is an on-request offering that coordinates, files, dissects and covers the impacts of online discussions happening crosswise over expert, customer produced and informal organization media destinations. Because of the knowledge gathered from this procedure, associations can comprehend the impacts online discussions are having on particular parts of their business activities [21]. Web based life examination (SMA) alludes to the approach of gathering information from online life destinations and writes and assessing that information to settle on business choices. This procedure goes past the standard checking or a fundamental examination of retweets or "preferences" to build up an inside and out thought of the social buyer. This is viewed as the essential establishment for empowering a ventures to:

- ✓ Execute centered commitment like balanced and one-to-numerous
- ✓ Enhance social joint effort over an assortment of business capacities, for example, client benefit, showcasing, bolster, and so on.
- ✓ Maximize the client encounter

With the growth of mobile technologies, the impact of social media is instant. It is useful in understanding customers in 3 important ways.

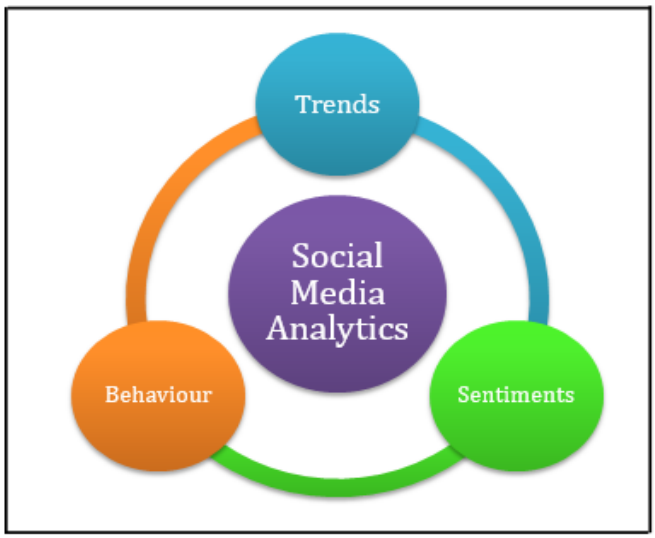

Figure 2 Social Media Analytics

Client slants are frequently helpful in understanding client feeling about the brand and its items and administrations. Internet based life showcasing is more about impact promoting. Online networking patterns are impermanent. These patterns are to a great extent impacted by financial, social, or political happenings. Henceforth, the clients" approach towards these patterns will likewise be periodical. In such a situation, advertisers can utilize these patterns to shape a procedure to build mindfulness about their items or administrations. Be that as it may, without the information the examination of genuine client conduct is unthinkable. This is the place internet based life examination come into the photo. The procedure of internet based life investigation starts by adjusting the accessible information to the business objectives. Advertisers need to utilize the information suitably to settle on more astute business choices [8, 9].

3. PROPOSED WORK

The proposed work is to identify and predict the social post popularity before posting content to a social media platform. In this context a data mining based data models are proposed and this section contains the details of the proposed approach.

3.1 System Overview

Data mining and their techniques are helpful for analyzing and understanding the hidden pattern in a large amount of data. Therefore the data models are prepared to establish relationships among the available attributes and according to the relationships among the attributes the future trends can be predicted. Now in these days the data mining and their techniques are also employed for the business analysis and their future possibility prediction. In this context the target domain data is analyzed using the data mining algorithms and using these algorithms the possibilities and future prospects of strategies are predicted.

The proposed work is intended to analyze the social media data for finding the future trends of a particular brand and their future growth. Therefore the social media post types and their user engagement in terms of like, share, comment and others are collected from facebook social page. Additionally for analyzing the trends the collected information is processed using the two different data mining algorithms i.e. clustering algorithm and a decision tree. The clustering algorithm is used for evaluation of post relevant data and creation of groups of social content according to their popularity. Here the popularity is clustered according to three major categories low, medium and high popular post. In addition of the decision tree algorithm is used for finding the two key outcomes first analysis of the historical data for establishing the relationship among the different post attributes. Additionally the second goal is invocation of the newly appeared content for finding the future marketing possibilities for the target content. This section provides a basic understanding of the proposed work. In next section the proposed model is explained in detail.

3.2 Methodology

The proposed data model is described in figure 3 in terms of the processes involved in the given system.

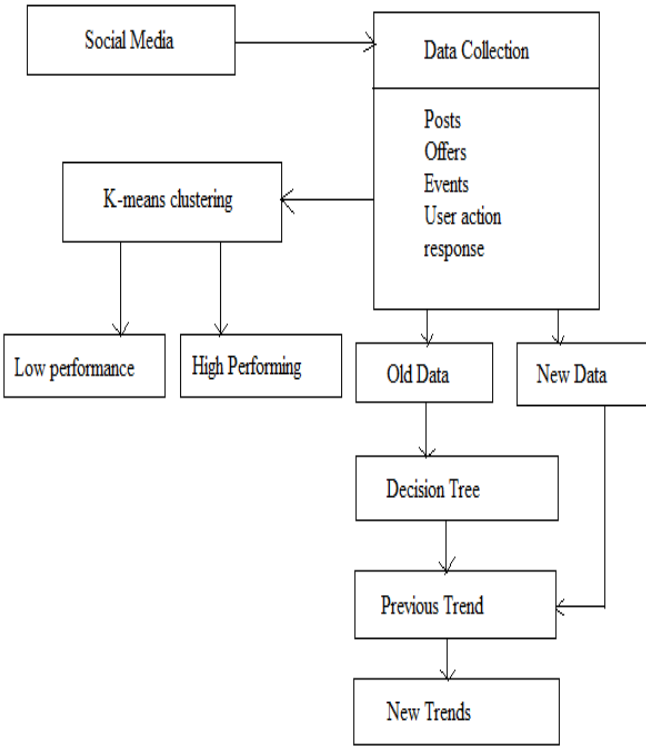

Figure 3 Proposed System Architecture

Social media data: Social media influenced various local and global factors of a business. Therefore different big and small scale companies are usages the social media for marketing and banding of their products. Therefore the different kinds of post (images, advertisements, videos and clip arts) are posted by the page managers, users in social media like, share and comment on these posts. But some kinds of post attract the peoples more and more therefore finding the popular trends of the type of post is helpful for increasing engagement of users for the particular product and brands. Therefore in this experiment the facebook social media targeted for post and user engagement analysis.

Data collection: That is a facebook post dataset which contains different attributes related to facebook post the dataset includes the following attributes:

Table 1 Dataset Attributes

S. No.	Attribute
1	Page total likes
2	Post Type
3	Category
4	Post Month, Post Weekday, Post Hour
5	Paid
6	Lifetime Post Total Reach
7	Lifetime Post Total Impressions
8	Lifetime Engaged Users
9	Lifetime Post Consumers
10	Lifetime Post Consumptions
11	Lifetime Post Impressions by people who have liked your Page
12	Lifetime Post reach by people who like your Page
13	Lifetime People who have liked your Page and engaged with your post
14	Comment
15	Like
16	Share
17	Total Interactions

K-means Clustering Algorithm: the collected data of social media posts in the format as given in table 3 is used with the classical k-means algorithm. The deployment of k-means algorithm is performed here for finding the high performance post and low performance post according to the dataset attributes. The traditional k-means algorithm is described as:

The K-Means clustering calculation is a parcel based group investigation strategy [10]. As indicated by the calculation we right off the bat select k questions as introductory bunch focuses, at that point compute the separation between each protest and each group focus and dole out it to the closest group, refresh the midpoints everything being equal, rehash this procedure until the point when the foundation work joined. Square mistake rule for clustering

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left\| x_{ij} - m_i \right\|^2$$

x_{ij} is the sample j of i-class, $m_i$is the center of i-class, n_ii is the number of samples of i-class. K-means clustering algorithm is simply described as:

Table 2 K-Means Algorithm

Input: N instances to cluster $(x_i,\ x_{i+1}\ \dots\ x_n)$, number of clusters k;
Output: k clusters and dissimilarity between each object and its nearest cluster center;
Process: 1. randomly select k instance as initial centroid $(m_1, m_2, \dots, m_k)$; 2. Calculate distance between each object Xi and each centroid, assign each object to nearest cluster, to calculate distance formula is given as: $d(x_i, m_i) = \sqrt{\sum_{j=1}^{d} \left(x_i - m_{j1}\right)^2}, i = 1 \dots N, j = 1 \dots k$ $d(x_i, m_i)$is distance between data i and centroid j. 3. Calculate mean of instances in each cluster as new centroid, $m_i = \frac{1}{N} \sum_{j-1}^{n_i} x_{ij}, i = 1,2, \dots, K$ N_i is number of samples of current cluster i; 4. Reiterate 2) 3) until the condition converged, and returns $(m_1, m_2, \dots, m_k)$

Low performance post: after processing of input data set using k-means clustering algorithm the entire dataset subdivided into two main group's namely low performance and high performance post. The low performance posts are those which have low popularity as compared to the high performance posts.

High performance post: the post which is most popular and has a significant amount like, share and comments.

Old data: that is the second part of data modeling for recovering or predicting future trends of a particular post. Therefore the collected dataset with all the attributes is considered as the old data or older post trends.

Decision tree: the decision tree algorithm consumes the data set and derives a tree data structure for demonstrating the relationship among the attributes and the class labels of datasets. In a dataset nodes are used for representing the attributes and the edges are used to denote the corresponding data values. Finally in leaf node the decisions are mounted on tree. Here for generating the decision tree model c4.5 algorithm is used. The c4.5 decision tree algorithm is reported as:

C4.5 (created by Quinlan, 1993) a calculation that takes in the choice tree classifiers, It has been watched that C4.5 performs short in the space where there is pre-passage of consistent qualities contrasted and the learning errands with for the most part isolate properties. For example, a framework which searches for all around characterized choice tree with 2 levels and after that put remarks [11]:

"The accuracy of trees made with T2 is equalized or even exceed trees of C4.5 upon 8 out of all the datasets, with the entire except one that have incessant attributes only."

INPUT: A data set D represent with discrete variables.

OUTPUT: A decision tree T which is constructed by passing data sets D.

1) A node (X) is created;
2) If instance falls in same class.
3) Make node (X) as leaf node and assign a label;
4) If attribute list is empty, THEN
5) Make node(X) a leaf node and assign a label of most frequent CLASS;
6) Choose an attribute which has highest information gain from list of attributes, and marked as test_attribute;
7) Confirming X in role of test_attribute;
8) To recognized value for every test_attribute for dividing samples;
9) Generating a fresh branch of tree that is suitable for test_attribute = att_i from node X;
10) Take an statement that Bi is a group of test_attribute=att_i in instances;
11) If Bi is NULL, THEN
12) Next, add a new leaf node, with label of most general class;
13) ELSE a leaf node is going to be added and returned by the decision_tree.

Previous trends: using the c4.5 decision tree algorithm a tree structure using the input social media data is generated. The traversing of the decision tree is used to define the previous data/post trends as leaf node of decision tree.

New data: the new data is a post which is needs to be evaluating using the decision tree and the possibility of the popularity among the social media.

New trends: using the new post the steps of publishing and possible popularity trends are predicted.

4. RESULT ANALYSIS

The given section includes the performance analysis of the implemented algorithms for the proposed data mining approach of social media analytics for brand building evolution. Therefore some essential performance parameters are obtained and listed with their obtained observations.

4.1 Space Complexity

The amount of main memory required to process the input data using the algorithm is known as memory consumption. Space Complexity of the system also termed as the Memory consumption in terms of algorithm performance. That can be calculated using the following formula:

$$\text{Memory Consumption} = \text{Total Memory} - \text{Free Memory}$$

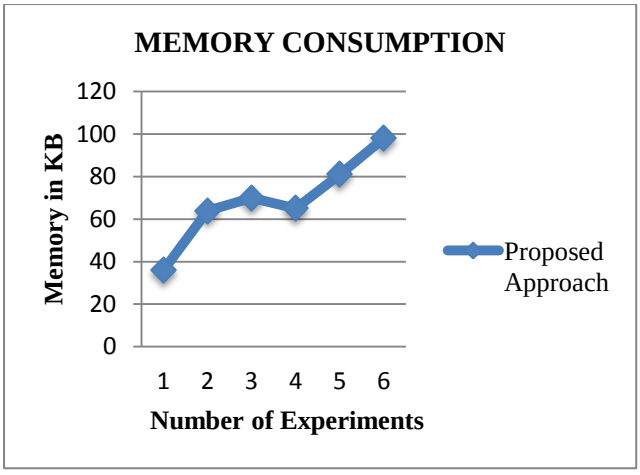

Figure 4 Space Complexity

The amount of memory consumption depends on the amount of data reside in the main memory, therefore that affect the computational cost of an algorithm execution. The performance of the implemented data mining approach of brand building is given using figure 4 and data is numerically show by table 3. For clarification of the result, X axis of figure contains the different amount of code execution and the Y axis shows the respective memory consumption during execution in terms of kilobytes (KB). According to the obtained results the performance of algorithm demonstrates similar behavior with input dataset. This consumed memory represents the required space by this algorithm process facebook dataset to produces efficient output.

Table 3 Numerical Values of Space Complexity

Number of Experiments	Proposed Approach
1	36
2	64
3	70
4	65
5	81
6	98

4.2 Time Complexity

The amount of time required to calculate future trend of the social media analysis using dataset is known as the time consumption of the system. That can be computed using the following formula:

$$\text{Time Consumed} = \text{End Time} - \text{Start Time}$$

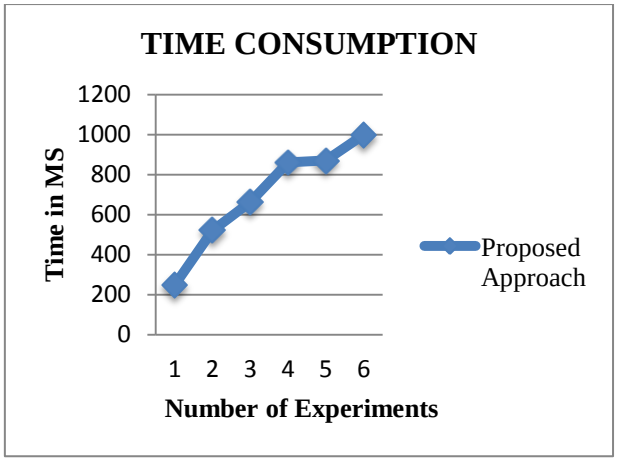

Figure 5 Time Complexity

The time consumption of the proposed algorithm is given using figure 5 and table 4. In this diagram the X axis contains the program execution of the system and the Y axis contains time consumed which is measures in milliseconds. Additionally, to demonstrate proposed we use blue line. According to the evaluated performance of the proposed technique is process the future tend by their post types. For processing algorithm consume time which is illustrated in table 4 in numerically.

Table 4 Numerical Values of Time Complexity

Number of Experiments	Proposed Approach
1	248
2	521
3	663
4	861
5	870
6	998

4.3 Accuracy

The performance of the correctly classified patterns using trending future post is represent in terms of accuracy. The performance evaluation of proposed approach is evaluated using classification concept. The accuracy of the data mining approach can be evaluated using the following formula:

$$\text{Accuracy} = \frac{\text{Total correctly classified Patterns}}{\text{Total input Words to Patterns}} \text{X100}$$

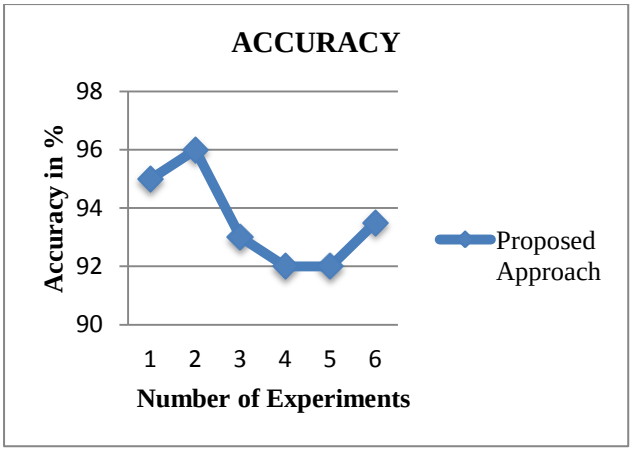

Figure 6 Accuracy

The accuracy of the implemented proposed algorithm of predicting social media analysis on brand building is represented using table 5 and figure 6. The given figure 6 contains the accuracy of the implemented algorithms. The X axis of the diagram shows the different experiments and Y axis contains the obtained performance in terms of (%). To demonstrate the performance of the proposed technique is representing using blue line and traditional approach is representing orange line. This technique is evaluated on the basis of input facebook dataset. According to the obtained results the performance of the proposed model provides more accurately recognized trending post. Additionally, the performance can be varying if we change the attribute of the input dataset.

Table 5 Numerical Values of Accuracy

Number of Experiments	Proposed Approach
1	95
2	96
3	93
4	92
5	92
6	93.5

4.4 Error Rate

The amount of data of misclassified patterns during classification of algorithms is known as error rate of the system. This can also be computed using the following formula.

$$\text{Error Rate \%} = \frac{\text{Total Misclassified Patterns}}{\text{Total Input Patterns}} \text{X100}$$

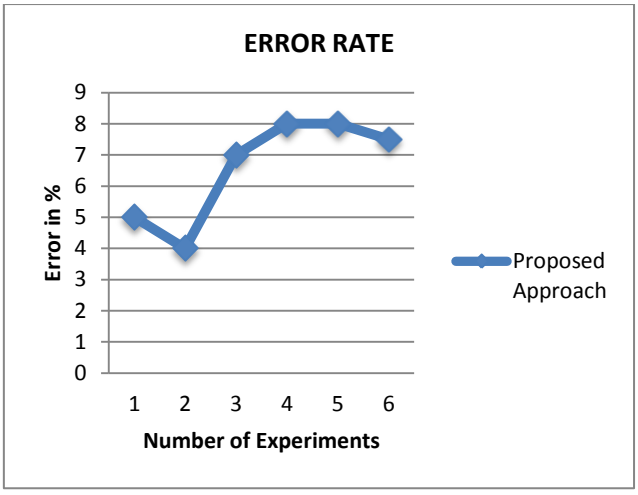

Figure 7 Error Rate

The figure 7 and table 6 shows the error rate of implemented algorithm of data mining. In order to show the performance of the system, X axis contains the experiments and the Y axis shows the performance in terms of error rate in percentage (%). The performance of the proposed data mining approach is given using the blue line. The performance of the proposed post classification is effective and efficient during different execution and reducing with the amount of data increases of future trend post. Thus the presented classifier C 4.5 is more efficient and accurate for classifying future trend post.

Table 6 Numerical Values of Error Rate

Number of Experiments	Proposed Approach
1	5
2	4
3	7
4	8
5	8
6	7.5

5. CONCLUSION AND FUTURE WORK

This section describes the summary of entire work performed for finding the future trends of the social media posts. The summary of work is given as the conclusion of work and future possible extensions are also explained in this section.

5.1 Conclusion

Now in these days for promotions and marketing social media is frequently used platform. The companies and individuals are post content on social media and according to user response on the post the popularity of product or brands are estimated. In this context if the posted data available in high quantity then the analysis of individual post and their trends are complicated task. Therefore the data mining and machine learning models are employed for finding the patterns or trends of the posted data. This analysis can be helpful for business intelligence and future product growth analysis. Therefore a model is proposed for analyzing the post data and obtains the two main objectives.

The main aim of the proposed system is two find two major outcomes:

1. Which kinds of post are performing effectively
2. What are the possibility to become a post popular

Therefore to compute both the objectives the two different data mining models are implemented namely k-means algorithm and C4.5 decision algorithm. The K-means clustering algorithm clusters the entire data in two categories high performing posts and low performing post. On the other hand the decision tree algorithm is used to analyze the historical post trends and based on the previous post's trends the new post popularity trends are predicted. Therefore first the collected historical post data is mounted over the decision tree structure and by traversing the developed decision tree possible trends of a new appeared post can be predicted using system.

The implementation of the proposed system is performed using the machine learning algorithms and JAVA technology. After implementation of the system performance is computed and the measured performance is summarized using table 6.

Table 5 Performance Summary

S. No.	Parameters	Remark
1	Time complexity	Low time complexity noticed for both the data model clustering and classification
2	Space complexity	Low space complexity is obtained for predicting the future content popularity

According to the obtained results as given in table 6.1 the proposed technique helpful for measuring the impact of social media post impact and identifying the possible future popularity growth of any kind of content marketing. Therefore the proposed model is acceptable for real world use with different applications.

5.2 Future Work

The main aim of the proposed work is to analyzing the social media post user engagement data for predicting the new post trends for any brand or product. The implementation of the required data mining technique is performed successfully. In near future the following extensions are proposed for work.

- ✓ Improving the decision making rules for achieving high accurate future trends of the posts
- ✓ Involving the text based analysis for estimating the users orientation for any kind of post
- ✓ Implementation of hierarchical clustering technique for estimating the reason of high and low performance posts

6. REFERENCES

[1] Altyeb Altaher, Ahmed Hamza Osman, "An Intelligent Approach for Predicting Social Media Impact on Brand Building", Journal of Theoretical and Applied Information Technology, Vol.95. No.17, 15th September 2017.

[2] Padma Janani M and Prabhalammbeka B S, "Predicting the Influences of Social Networks on Brand Building", International Journal of Mechanical Engineering and Technology (IJMET) Volume 8, Issue 10, October 2017, pp. 308–317.

[3] Bernardo Varela Vala, "The impact of Social Media in Brand Building", MS Dissertation, ISCTE Business School, September 2015

[4] Shen zihao, Wang Hui, "Research on E-commerce Application Based on Web Mining", 2010, IEEE.

[5] Jaideep Srivastava, Prasanna Desikan and Vipin Kumar, "Chapter 21- Web Mining — Concepts, Applications, and ResearchDirections",

http://dmr.cs.umn.edu/Papers/P2004_4.pdf

[6] "Chapter 1: Introduction", Sodhganga, available online at: http://shodhganga.inflibnet.ac.in/bitstream/10603/11075/5/05_chapter1.pdf

[7] Web Content Mining: Tutorial given at WWW-2005 and WISE-2005 available online at: https://www.cs.uic.edu/~liub/WebContentMining.html

[8] Yogeswari Suppiah, Raja Mohd Tariqi Raja and Mohd Fahmi Mohamad Amran, "A Study on Social Data Analytics and Privacy Concern among Social Media Users", International Journal of Computer Applications (IJCA) Volume 149 – No.5, September 2016.

[9] "Social Media Analytics: Driving Better Marketing Decisions: Insights into Customer Behavior", White paper, Cybage.

[10] Wang, Juntao, and Xiaolong Su. "An improved K-Means clustering algorithm." In Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on, pp. 44-46. IEEE, 2011.

[11] Mishra, Kundan Kumar, and Rahul Kaul. "Audit Trail Based on Process Mining and Log." International Journal of Recent Development in Engineering and Technology 1, no. 1 (2013).