Marketing Tactics Improvement by Looking out the Key Users from Facebook

Bhushan Talekar
M.E – Computer Engineering
VIVA Institute of Technology

Umesh Mohite
B.E – Computer Engineering
VIVA Institute of Technology

Pragati Patil
M.Tech – Information Technology
VIVA Institute of Technology

Pallavi Vartak
M.E – Information Technology
VIVA Institute of Technology

ABSTRACT
Talking about running a lasting business, a social media presence is critical. Understanding the interest of all users & based on it, publishing the required information as per their tastes is an important factor, when it comes to establishing a social media presence that makes an impact. For advertising campaig, discovering the appropriate target markets and audience is an important stage in the market research.

Identifying the target users, Designing of market strategy/plan, Building the marketing network (groups) & Statistical analysis of categories are the four important tasks we aim to focus on. Categories have been found based on their influence by using clustering technique. Further this paper helps to extract emotional feelings of the user so that any related articles, posts or videos can be posted to that user.

Keywords
Clustering, FCM, K means, Facebook Graph Api.

1. INTRODUCTION
For the last few years, social media phenomenon emergence has been one of the most remarkable developments in the world of Internet. Internet is the powerful tool and the ability to connect with people around the world.

Social media lets people communicate either directly or via media objects. The year 2006 can be regarded as the break-through year of social media. At that point, the popular early applications like Wikipedia and MySpace had gathered significant numbers of users, while Facebook and YouTube had been introduced to the public. YouTube since early 2006 and Facebook since early 2007 after it opened its doors to anybody to register. Our work is concentrating mainly on Facebook.

2. PROBLEM DEFINITION
Quality improvement is an important factor for any business. But, the question is that how to move the users towards our product? & how to find who is interested in knowing our new products, versions, features, facilities etc.

Social media is used to find the users. Here we have proposed a competent design & a clustering technique to grow up the advertising way towards identifying the key users using Facebook.

3. METHODOLOGY
Fig.1 shows the steps involved in applying clustering algorithm to find the key users i.e.

1. **Preprocessing:** Includes Training of the system.
2. **Extraction:** Includes Data extraction from Facebook.
3. **Filtering:** Includes Tokenization and Cleaning functions.
4. **Clustering:** Includes Classification of Post Message andComment into different categories
5. **Identifying targeted users:** Includes finding of influential users.
6. **Design of Marketing strategy:** Put out the new posts to key users based on their interests.

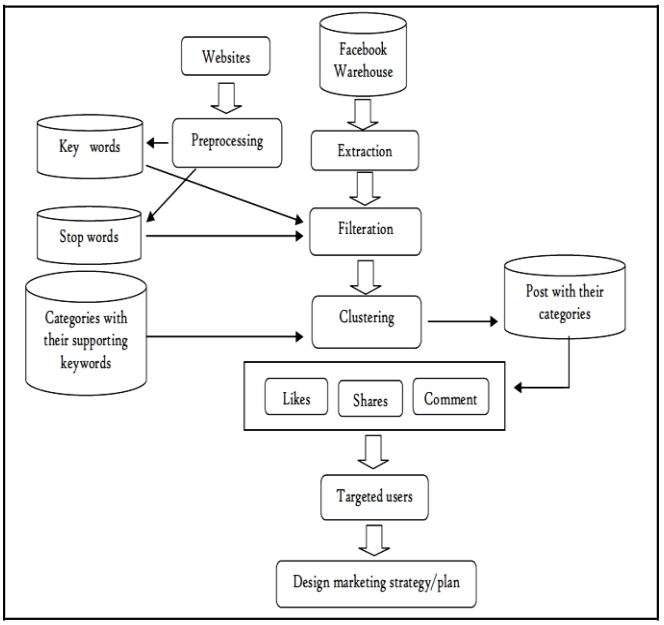

Fig.1 System Architecture

1. Preprocessing:
Includes training of the system. Targeting potential users on Facebook is not so easy with social networking sites. Without permission nobody can access the user's profiles, but on the fan page we can promote businesses. For good analysis of post we have collected thousands of keywords also called buzz words related to posts'. It is important to select the significant keywords that carry the meaning, and discard the words that do not contribute to distinguishing between the posts. Keywords are defined as a sequence of one or more words and provide a compact description of a post's content.

2. Extraction:
In the Extraction we include Data extraction from Facebook. A basic token is assigned to any Facebook user and can be used to search any publicly available information. This will still not include a "friends" list. However, we can search for wall posts for any known Facebook IDs as long as we have a basic authentication token and the Facebook user has declined to make this information private. This will allow us to see when people in our list have posted to each other's walls for a given time range. Facebook allows you to search based on an ID. Basic format for sending HTTP requests to the Facebook API also known as the Graph API is detailed in [6] & [7] for an example of this approach.

3. Filtering:
As the extracted information are stored in a separate database as knowledge source, which is retrieved for further tasks. Filtering includes Tokenization and Cleaning functions. Filtering is done based on the list of stop words and stemming words, which are analyzed or mined by examination from the training corpus. Stop-words, which are language-specific functional words, are frequent words that carry no information.

Stemming techniques are used to find out the root/stem of a word. Stemming converts words to their stems, which incorporates a great deal of language dependent linguistic knowledge .These stop words and stem words are stored in StopWords table and StemWords table of databases respectively in the preprocessing phase for providing assistance in the process of filtering the contents of message. A sample of collected stop words and stem words are shown in Table I and Table II. [8][9][10]

Table I. Sample – Stop Words List

able, about, above, according, after, afterwards, again, ain't,… be, became, because, behind, being, brief, but, by,… came, can, can't,… each, edu, eg, etc, even, ever,… from, further, furthermore,… had, hadn't, happens, hardly, has, hasn't, have, he, hello, help,… her, here, hereby, his, indeed…….

Table 2. Sample – Stemming Words List

Word	Stem	Word	Stem
Consign	consign	Consist	consist
Consigned		Consisting	
Consigning		Consisted	
Consignment		Consists	

We have used the Porter Stemmer algorithm to perform the stemming of keywords into stem words. [8][9]

Algorithm: Porter Stemmer Algorithm

1. Gets rid of plurals and -ed or -ing suffixes
2. Turns terminal y to i when there is another vowel in the stem
3. Maps double suffixes to single ones: -inaction, - national, etc.
4. Deals with suffixes, -full, -ness etc.
5. Takes off -ant, -ence, etc.
6. Removes a final –e

4. Clustering
Clustering partitions the data set into clusters or equivalence classes. We have clustered users into different categories based on the posts made by them on Facebook. For comparative analysis of the results, we have made use of two clustering algorithms, Fuzzy C-Means & K-Means respectively.

i. Fuzzy C-Means Algorithm
Fuzzy C-means (FCM) is a method of clustering which allows one piece of data belong to two or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition.

The FCM algorithm is composed of the following steps:

Algorithm: Fuzzy C-means algorithm

1. Let us suppose that M-dimensional N data points represented by xi here (i-1,2,…N), are to be clustered
2. Assume the number of clusters to be made, that is, C, where

$2 \leq C \leq N.$

3. Choose an appropriate level of cluster fuzziness f >1.
4. Initialize the N * C * M sized membership matrix U, at random, such that Uij ϵ [0, 1] and for each I and fixed value of m.
5. Determine the cluster centers Cjm, for the jth cluster and its mth dimension by using the expression given below:
6. Calculate the Euclidean distance between ith data point and jth cluster center with respect to, say mth dimension like the following:

$$D_{ij} = \sum_{1}^{n} \sqrt{|x_i - C_j|^2} \qquad --(ii)$$

7. Update fuzzy membership matrix U according to Dik, If Dik > 0, then

$$U_{ik} = \frac{1}{\sum_{j=1}^{c} (\frac{D_{ik}}{D_{jk}})^{\frac{2}{m-1}}}, 1 \leq i \leq c, \quad 1 \leq k \leq N \qquad --(iii)$$

If $D_{ik} = 0$, then the data point coincides with the corresponding data point of j^{th} cluster center C_j and it has the full membership value, that is, $U_{ij} = 1.0$

8. Repeat from step 5 to step 7 until the changes in U $<^{\varepsilon}$, where ε is a pre-specified termination criterion.

ii. K-Means Algorithm
K-Means algorithm is a hard partitioned clustering algorithm widely used due to its simplicity and speed. It uses Euclidean distance as the similarity measure. Hard clustering means that an item in a data set can belong to one and only one cluster at a time. It is a clustering analysis algorithm that groups items based on their feature values into K disjoint clusters such that the items in the same cluster have similar attributes and those in different clusters have different attributes.

The K-means algorithm is composed of the following steps:

Algorithm: K-means algorithm

1. Define the number of clusters K.
2. Initialize the K cluster centroids. This can be done by randomly selecting K data items from the data set.
3. Calculate the Euclidean distance between ith data point and jth cluster center, like the following:
4. Assign each item to the cluster with the nearest centroid. In this way all the items will be assigned to different clusters such that each cluster will have items with similar attributes.
5. After all the items have been assigned to different clusters re-calculate the means of modified clusters by taking the average coordinate among the items. The newly calculated mean is assigned as the new centroid.
6. Repeat step (iii) until the cluster centroids do not change.

iii. Illustration of Fuzzy c-means and K-means using sample data

Here, we will consider 2 categories mainly- Technology as cluster-1 & Sports as cluster-2.

Firstly, we generate a domain specific data set for all objects i.e. posts. In the next phase the post contents is divided into tokens & non-significant characters are removed (hyphens, stop words, white spaces, tags are removed).Tokenization & Cleaning is done in this phase. In this phase, calculation of Initial Matrix is done by analyzing the FCM algorithm & K- means algorithm to place the posts into clusters with different instances.

Table 3. No. Of Occurrences Of Two Categories

Special Words	Occurrences	Occurrences
Mobile	100.201	5.234
Game	120.345	150.456
Cricket	2.223	201.432
Job	10.562	15.785

We have taken the sample posts that are related to the categories of Technology & Sports respectively.

Now we find out the number of occurrences of the special words (which helps to cluster posts into categories) in each posts respectively. Table III. gives a clear idea of the occurrences.

Seeing a table, we can identify that the "Game" & "Job" belong to both groups with equal amounts. Whereas the words like Mobile & Cricket have more differences between the categories i.e. Mobile (100,5), Cricket(2,201).

1	0.87	0.03	0.13
0	0.13	0.97	0.87

Let's consider these 2 as special words for further processing, the next phase is clustering wherein we are going to cluster the given posts & classify them to categories.

Let M= {m1, m2, m3, m4…mn} represents m messages (posts content) to be clustered.

Each of these messages, mi, is defined by the,'s' special words i.e. mi= {mi1, mi2, mi3…….mis}

Here, each mi in the universe M is a s-dimensional vector representing the "s" special words which will be normalized using equation below:

	Post 1	**Post 2**	**Post 3**	**Post 4**
Mobile	1	1	4	5
Cricket	1	2	3	4
RESULT	**Post 1 Belongs**	**Post 2 Belongs**	**Post 3 Belongs**	**Post 4 Belongs**

Occurrences=WC/TC *100

WC-Word Count

TC- Total no. of words in a message

Consider we have 4 posts collected from Facebook. The table below shows the sample occurrences of special words in these 4 posts.

	Post 1	**Post 2**	**Post 3**	**Post 4**
Mobile	1	1	4	5
Cricket	1	2	3	4

Now we perform Fuzzy C-means & K-means one by one ,to cluster posts into required categories (Technology & Sports). Initial parameters:

a. No. of clusters, k = 2 (Technology & Sports)
b. Fuzziness component, m = 2
c. Termination criterion, £ = 0.03

A. By Fuzzy C-Means Algorithm:

1. Initial Cluster Center

Assume post 1 & post 2 as initial cluster center. So, C1 = (1,1) & C2 = (1,2)

2. Euclidean Distance Matrix (D0)

0	1	3.61	5
1	0	3.16	4.47

3. Initial Matrix G0

1	0	0	0
0	1	1	1

4. New Cluster Center

C_1 = (1,1) & C_2 = (3.33,3)

5. Euclidean Distance Matrix (D1)

0	1	3.61	5
3.07	2.54	0.67	1.94

6. Update Matrix G1
After multiple iterations, we get the final updated matrix as follows:-

Final Updated Matrix G

0.99	0.98	0.04	0.02
0.01	0.02	0.96	0.98

Termination Criterion met. So, we stop.

RESULTS
Post 1 & Post 2 have higher occurrence keyword "Cricket" which belongs to the SPORTS category, so we conclude that both posts belong to the SPORTS category.

Post 3 & Post 4 have higher occurrence keyword "Mobile" which belongs to the TECHNOLOGY category, so we conclude that both posts belong to the TECHNOLOGY category.

Now, we shall solve the same example by K-means algorithm, as follows:-

B. By K-means algorithm:

1. Initial Cluster Center
Assume, post 1 & post 2 as initial cluster center. So, C1 = (1,1) & C2 = (1,2)

2. Euclidean Distance Matrix (D0)

0	1	3.61	5
1	0	3.16	4.47

3. Initial Matrix G0

1	0	0	0
0	1	1	1

4. New Cluster Center
C1 = (1,1) & C2 = (3.33,3)

5. Euclidean Distance Matrix (D1)

0	1	3.61	5
3.07	2.54	0.67	1.94

6. Update Matrix G1

1	1	0	0
0	0	1	1

7. New Cluster Center

C1 = (1,1.5) & C2 = (4.5,3.5)

8. Euclidean Distance Matrix (D2)

0.5	0.5	3.35	4.72
4.30	3.81	0.71	0.71

9. Update Matrix G2

1	1	0	0
0	0	1	1

Here, $G^1 = G^2$, groups don't change anymore. So we stop.

RESULTS

	Post 1	**Post 2**	**Post 3**	**Post 4**
Mobile	1	1	4	5
Cricket	1	2	3	4
RESULT	**Post 1 Belongs To**	**Post 2 Belongs To**	**Post 3 Belongs To**	**Post 4 Belongs To**

Post 1 & Post 2 have higher occurrence keyword "Cricket" which belongs to the SPORTS category, so we conclude that both posts belong to the SPORTS category.

Post 3 & Post 4 have higher occurrence keyword "Mobile" which belongs to the TECHNOLOGY category, so we conclude that both posts belong to the TECHNOLOGY category.

5. Identification of Targeted Users:
Identify the people who are interested in information related to a particular category like Advertisement, Sports, Politics, Entertainment, Social awareness, etc.

The process uses selection of category to identify the interested users from the posts. The process involves Extraction of users who have liked or shared or commented the post/s in that category. Prepare the database of users interested in each category.

6. Design of Market Strategy:
To promote the new post to a set of users, first find the category of the post. Then select the targeted users of that post. Thus, Multicast the post to only interested users.

RESULTS
From the above solved example & other test results, it is very clear that both the clustering algorithms generate similar clusters & hence similar results. But, what differs is the time efficiency between the two algorithms. Figure 2, shows the time efficiency of Fuzzy C-Means algorithm vs K-Means algorithm. It is very evident from the figure that K-Means algorithm has a better Time efficiency as compared to the Fuzzy C-Means algorithm. Fuzzy C-Means requires more computation time than K-Means because of the fuzzy measures calculations involvement in the algorithm. Figure 3 & Figure 4 show the clustering of 5 users on Facebook. Users are clustered on the basis of the LIKES made by them on Facebook & on the basis of the POSTS written by them on their Facebook wall.

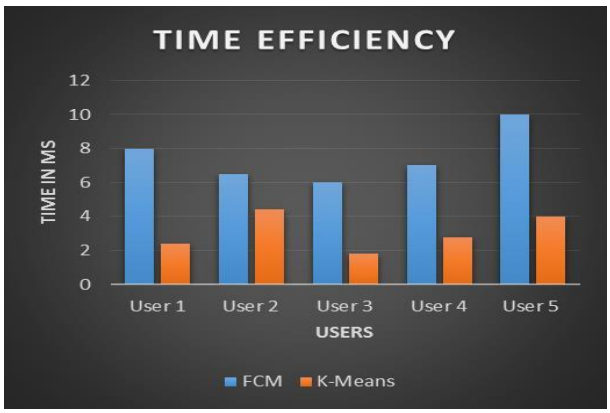

Fig.2 Time Efficiency of FCM & K-Means algorithm

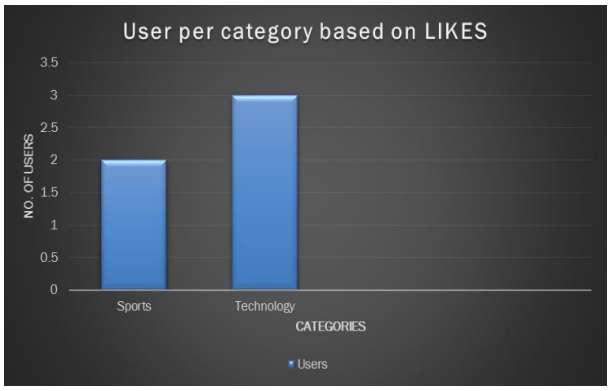

Fig.3 Number of Clustered Users per category based on Likes

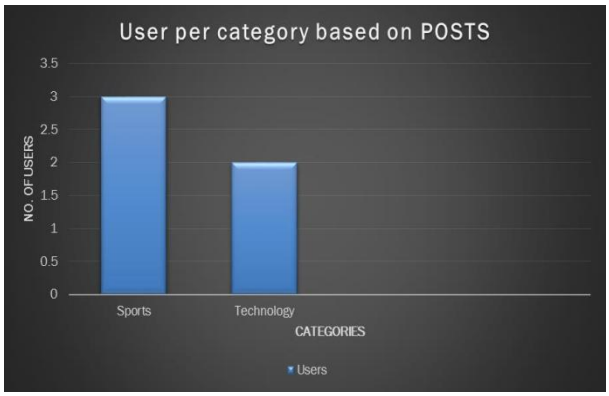

Fig.4 Number of Clustered Users per category based on Posts

4. CONCLUSION

This research successfully classifies the Facebook users into specific category based on their data from the Facebook. The application makes the right use of Facebook Graph Api, to fetch user's personal information with his permission & allowing the application to perform clustering algorithms on the user's data to classify him into a specific category.

The research can be a great boost to the business world. It can help businesses, market their products & services to the targeted customers only. This will surely improve their marketing tactics & help them reach to their desired group of customers directly. As Facebook is the most used social media in today's world, this research will definitely be a boom to all the businesses who wish to create their identity in the online marketplace.

Here, the experimental results are promising: Performance of Facebook Graph Api is simply amazing. It ensures to fetch all the user's desired data within few steps. The performance of K- Means algorithm is better compared to the Fuzzy C-Means algorithm. Hence, classification is faster and more efficient.

5. REFERENCES

[1] Toni Ahlqvist, Asta, Minna Halonen & Sirkka Heinonen "Social Media Roadmaps Exploring the futures triggered by social media", VTT TIEDOTTEITA- Research Notes2454, ESPOO 2008.

[2] Assaad, Waad; Jorge Marx Gomez. Social Network in marketing (Social Media Marketing) Opportunities and Risks 2 (1). Retrieved 7 February 2013.

[3] Jessica Bosari "The developing role of social media in the modern business world", http://www.forbes.com/sites/mone ywisewomen/2012/08/08/the-developing-role-of-social-media-in-the-modern-business-world/.

[4] Improving Revenue and Customer Engagement With SocialMedia Analytics" –A retail Touchpoints White Paper,sponsored by SAS, 2001. James Carson "Whats the score? Ultimate guide to social scoring" 2013, http://www.fliptop.com/socialscore/

[5] Gutwin C, Paynter G, Witten I, Nevill-Manning C and Frank E. "Improving browsing in digital libraries with keyphrase indexes", Decision Support Systems 27(1–2), 81–104, 1999

[6] Facebook Graph Api details : http://developers.facebook.com/docs/reference/api/

[7] http://zesty.ca/facebook/

[8] Media Analytics" –A retail Touchpoints White Paper,http://snowball.tartarus.org/algorithms/porter/stemmer.html

[9] "Overview of Stemming Algorithms" Ilia Smirnov http://the-smirnovs.org/info/stemming.pdf

[10] Anjali Ganesh Jivani et al, Int. J. Comp. Tech. Appl., Vol2 (6), 1930-1938 – "A Comparative Study of Stemming Algorithms ".

[11] Mitchell D'silva, Deepali Vora / International Journal ofEngineering Research and Applications (IJERA) ISSN:2248-9622 www.ijera.com Vol. 3, Issue 1, January - February 2013, pp.1267-1275 – "Comparative Study of Data Mining Techniques"

[12] Sumit Goswami and Mayank Singh Shishodia/International Journal of Data Mining & KnowledgeManagement Process (IJDKP) ISSN : 2230 -9608[Online] ; 2231 - 007X [Print] – "A Fuzzy Based Approach To Text Mining And Document Clustering"

[13] Soumi Ghosh, Sanjay Dubey , "Comparative Analysis of K-Means and Fuzzy C-MeansAlgorithms",((IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 4, No.4, 2013

[14] Karuna C.Gull,Akshata Angadi,Seema C.G,Suvarna G.Kanakaraddi, "A clustering technique to rise up the marketing tactics by looking out thekey users" 2014IEEE International Advance ComputingConference(IACC)