Evaluating Conversion Rate from Advertising in Social Media using Big Data Clustering

Khaled H. Alyoubi[1], Fahd S. Alotaibi[1]

[1]Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21541, Saudi Arabia

Abstract
The objective is to recognize the better opportunities from targeted reveal advertising, to show a banner ad to the consumer of online who is most expected to obtain a preferred action like signing up for a newsletter or buying a product. Discovering the most excellent commercial impression, it means the chance to exhibit an advertisement to a consumer needs the capability to calculate the probability that the consumer who perceives the advertisement on the users browser will acquire an accomplishment, that is the consumer will convert. On the other hand, conversion possibility assessment is a demanding process since there is tremendous data growth across different information dimensions and the adaptation event occurs infrequently. Retailers and manufacturers extensively employ the retail services from internet as part of a multichannel distribution and promotion strategy. The rate at which web site visitors transfer to consumers is low for online retail, out coming in high customer acquisition expenses. Approximately 96 percent of web site users concluded exclusive of no shopper purchase[1].This category of conversion rate is collected from the advertising of social media sites and pages that dataset must be estimating and assessing with the concept of big data clustering, which is used to group the particular age group of people along with their behavior. This makes to identify the proper consumer of the production which leads to improve the profitability of the concern.

Key words: *Social Media, Big Data Clustering, Conversion Rate, Evaluation,*

1. Introduction

Day to day the social network account holders of Instagram bring out approximately 50,000 photos. Twitter account holders tweet 473,400 updates as well as 4.3 million viewers view the YouTube to watch a video content [1]. Most of those counts and activities will matter for the brand. Discovering significant social network media trends, signals as well as cues recently requires broad analytics potentials.

Social media afford overwhelming amounts of dynamic information. Extracting knowledge from these dimensions needs mechanization. Computing promptly in excess of this statistics is a dispute for both techniques and structural designs. Recently, micro-blogging has become an admiredfashion which is dependable for a great amount of information distribution. The most ubiquitous micro-blogging service is Twitter, which is a fashionable tool for undersized, recurrent communication. Twitter account holders tweet about different themes or positions as well as they follow different accounts with different perspectives. Any person around the globe can employ Twitter to speak about each day behaviors and search for information.

Monitoring such elevatedprocedure of Twitter has recommended various businesses and research groups to investigate these tweets to estimate potential connections as well as predictions [13]. Several tools and applications have been enhanced around this view to predict outlook events. As the consequence of Twitter climbs, the linked research in the fields of corporate, information science, along with civil society sectors is flourishing.

Social media has made with social networks ubiquitous moreover given researchers access to substantial amounts of data for observed analysis. These data base present a wealthy source of confirmation for learning energetic of human being along with group behavior, the formation of networks and universal patterns of the flow of information. For example, Facebook consists of more than 400 million active users sharing over 5 billion quantitiesof information every month. The micro-blogging services such as Twitter are causing a worldwide fashion [2].

Manuscript received July 5, 2021
Manuscript revised July 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.7.35

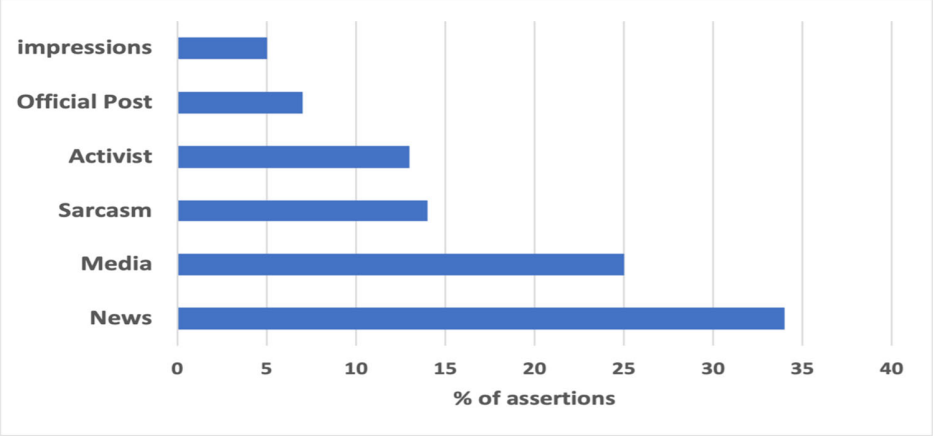


Fig 1 Analyzing Tweets for excellence in Journalism[2]

Greater than 280 million monthly active holders send more than 500 million comments for every day and 80% of those lively users send tweets from mobile. At every moment of time 1.45 million twitter queries are generated, the activity of searching content is one of the most admired events of Twitter. Further the upcoming popular activity is time-line presentation. There are 1.22 million time-line views each minute. The size of tweets is massive with more than 347,000,140 characters are going out in the form of messages in each 60 seconds. Tweet comments can be classified into trends and are interrelated with tags and followers of social associations. The classification is neither perfect nor successful due to the diminutive length of tweet messages along with noisy data corpus [15].

Figure 1 is one of the sample instances of analyzing tweeter comments for excellence in Journalism which demonstrates the percentage ratio of news & information, photos & videos, jokes, hopes & prayers for safety, political commentary as well as excitement. This information are collected from the tweeter and categorized into various formations with the percentage level which is portrays in the graph.

II. CONVERSION RATE

Conversion rate refers to the percentage of page or site visitors of the specific websites. It means the proportion of people or concerns that shift from the specific stage to the further stage in a process [3]. General example embrace the percentage of human mind who visit the web site as well as sigh up and also some of the people get a trial of the products and end up with purchasing. For sales and business promotion, the primary objective of estimating conversion rates is to discover week spots in a concern's customer acquirement process. This task is acknowledged as conversion optimization [16]. Conversion rating is also utilized as a metric for evaluating the behavior of promotion initiating scheme. Several structures are employed when investigating the ratios. This can also included with digital funnel marketing, brand health, sales funnel and the priority of effects. Conversion rate is measured by the number of conversions divided through the number of visitors referring the tweet.

$$Cr = (Cv/Tv) * 100\ \% \quad (1)$$

Where,

Cr = Conversion Rate
Cv = Converting Visitors
Tv = Total Visitors

Figure 1 is one of the sample instances of analyzing tweeter comments for excellence in Journalism which demonstrates the percentage ratio of news & information, photos & videos, jokes, hopes & prayers for safety, political commentary as well as excitement. This information are collected from the tweeter and categorized into various formations with the percentage level which is portrays in the graph.

Using the formula (1) Conversion Rate is calculated with the support of converting visitors and total visitors [3]. For instance, the particular web site had 17,492 visitors and in the last month 2,305 conversions. Then the conversion rate is calculated as 13.18%.

GOOD CONVERSION RATE

The rate of conversation is fluctuates noticeably depending on the business, industry, quality of the traffic, selling product as well as specific action of the tracking conversion. As an outcome, to discover broad conversion

rate statistics out there, the qualifies as a superior conversion rate for a specific business and the marketing campaign. A conversion is not constantly the similar thing as a purchase. The conversion rate is a usable metric; the objective of the major marketing is not to produce conversions [14].

Consider the grocery store's *brand health* data in figure 2 has to be measured with the support of conversion rate calculations. Through this calculations the factors similar to aware, visited last 12 months, visited last month, visited last trip along with loyal are considered. Initially conversion rate from the factor of awareness to visit in the last 12 months is estimated as (93/99) 94%. For visited last month relative to visited in last 12 months is (81/93) 87% and so on. The big weak mark evident through the analysis of conversion rate is that loyalty of the customer is a challenging factor due to the conversion rate 19% of those who visited being loyal. The concern can understand that they need to give more preference of the factor loyal presumably they visited due to convenience or a price discount of some kind.

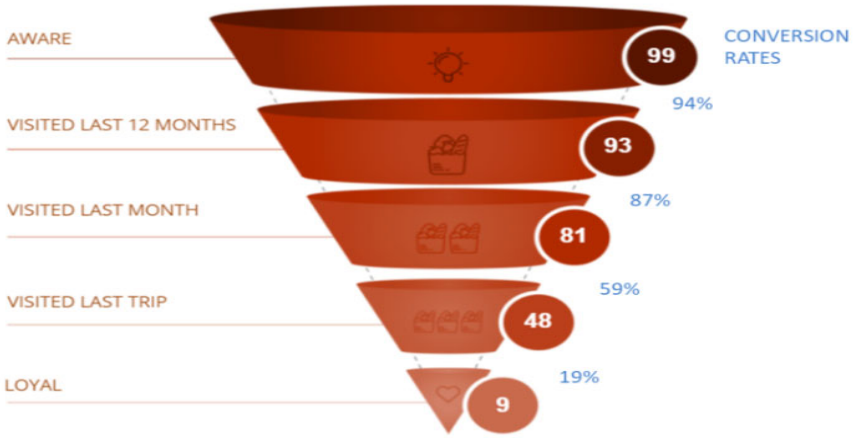


Figure 2 Sample Conversion Rates of Brand Health Data

CONVERSION RATE FRAMEWORKS

Most of the frameworks are extensively utilized when investigating conversion rates. The trend of digital marketing, there are countless alternatives of a *funnel, it means* the diagrammatic representation of the current business scenario. They illustrate various stages of the attainment procedure of the business like impressions through ads, social media, visiting the home page, searching, adding a product to a basket, going to checkout, purchasing, and regularly purchasing [4].

In trade, funnels are mostly utilized but tend to be described brand health metrics. They illustrate an assortment of activities from consciousness through there liability. In recent market with profound amounts of brand advertising, the *hierarchy of effects* model is also universal, tracking conversion through various stages of awareness, knowledge, consideration, preference, purchase, and recommendation.

III. BIG DATA CLUSTERING

Clustering is an unsupervised learning task where to be aware of a limited understanding of classifications named clusters to describe the data. Clustering is similarly exemplifies as a gathering of the comparable elements accumulated tightly collectively. Clustering separates the data sets of population to dissimilar clusters based on the homogeneity of dissimilar classes. Various metrics are utilized by unusual clustering algorithm to achieve clustering of data [5]. For instance, figure 3demonstrates several real time clusters. Several incoming entity will be classified in to diverse groups based on the extracted features of the given entity to organize clusters. The grouping progression utilizes the cluster qualities of inter and intra class similarities of different entities to construct the final pronouncement to cluster groups. Anoptimized Clustering technique will generate top quality clusters with high intra-class comparability - Similar to each other inside a similar group low between class likeness - Different to the objects in different groups [13].

Figure 3 portrays the samples of social media clustering. The social media datasets of news and politics,

Sports, soccer along with entertainment are gathered and applied with the technique of clustering. Then each cluster is segregated according to their similarity, for example the entertainment dataset is divided into movie, music etc.,

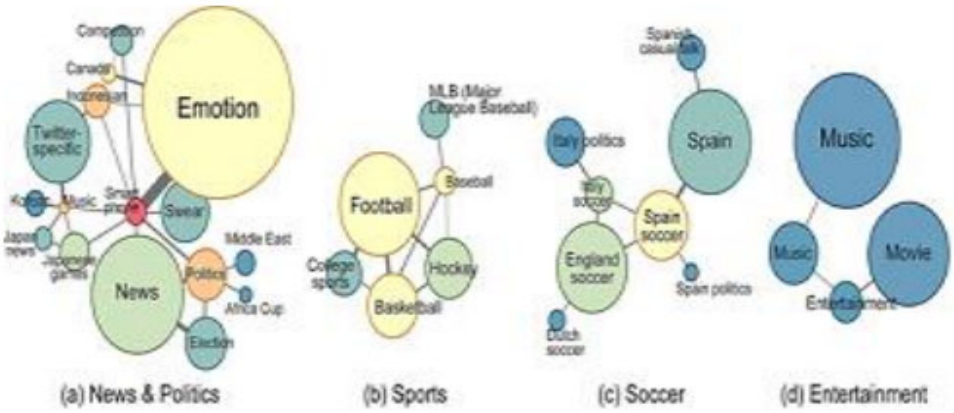


Figure 3 Sample dataset of Social media with Big Data Clustering

IV. CANOPY CLUSTERING

Canopy clustering is a rapid and fairly accurate clustering algorithm. It separates the key in data elements into overlapping groups called canopies. Two various distance threshold values are utilized for the evaluation of the cluster centroids. Canopy clustering can offer a speedy approximation of the number of clusters and preliminary cluster centroids of a given dataset. It is essentially utilized to recognize the data and afford input to techniques like k-means [17]. Normally, canopy clustering is employed for the purpose of grouping the data points into spherical regions but these regions cannot be overlap, and most of the data elements are allocates too many regions.The objective of the Canopy clustering is for selecting initial centroids more resourcefully than randomly, particularly for the real time applications of big data. All data elements are represented as a specific point in a feature space of multidimensional. The technique employs a rapid estimated distance metric and two threshold distance values T1 > T2 for processing. The essential algorithm is to commence with a set of data elements and eliminate randomly. Generate a Canopy holds this data object and generate with the help of the remainingdata objects. At every data element, whether its space from the initial data object is < T1, then include the data element to the relevant cluster points. Additionally, if the space is < T2, then eliminate the data element from the entire dataset. Accurate relevant data points are discovered and identified to that appropriate cluster like this way the data elements are classified into the specific groups. The iteration of the clustering loops completed in anticipation of the beginning dataset is vacant;gather a set of canopy data elements, every group enclosing one or more number of data objects. A specified data object may happen in more than one Canopy centroid [6].

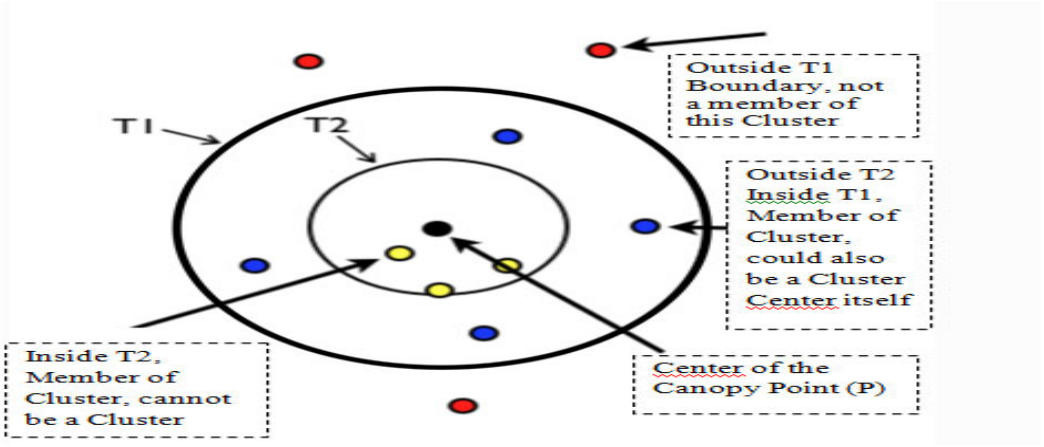


Figure 4 Canopy Clustering Technique

Figure 4 portrays the concept of Canopy Clustering Technique. In the above figure is filled with more number of data elements with various colours like black, yellow, blue and red. To describe the proximity region is referred as Canopy and to draw a circle or hypersphere centered at a data point. The data elements formed outside this sphere is considered to be far away. The black data element is considered as a centroid data element as well as it is referred as center of the Canopy (P). According to the nearest value of remaining data elements it will formulate the different groups with the support of the two threshold values are fixed as T1 and T2. In the above said figure the yellow coloured data elements are formed in between the canopy center point and within the radius threshold value T2, these points are very close and not approved to start a new canopy clustering process. Moreover, the blue coloured data elements are formed outside T2 and within radius of T1, as well as these points are close enough to be the members of this cluster, but they can eligible to start a new canopy clustering process. And the final possibility of red coloured data points are considered as too far away and positioned in outside the threshold boundary of T1.

ALGORITHM OF CANOPY CLUSTERING

STEP 2 : Among the threshold values T1 > T2 can be determined through cross checking.

STEP 3 : Consider any data point D from the dataset, and rapidly measure the distance among the data points D and entire canopy with a low estimation cost method.

STEP 4 : If there is no canopy presently, utilize data point D as a Canopy.

STEP 5: If data point D and a canopy, the space is within T1 then include the data element D to this canopy cluster.

STEP 6 : If the distance among the data point D and a specific canopy is within T2, it require to delete point D from the dataset list.

STEP 7 : Whether the data point D is very close enough to the canopy, then the centre of canopy is declared.

STEP 8 : Repeat the steps from 2 to 6 until the data sets gets empty and ends.

V. RESULT ANALYSIS

New trends precious the media production recently, adapt the essential characteristics and approaches of companies and consumers as well. Various firms consider that their occurrence on the surface of social media is the key factor to success. Social networks have become an element of regular life for most of the mankind around the globe today. A social media is a wonderful use case for bottom up design. This permits future promotions to be greatly non-breaking while being comparatively simple to enhance and sustain. This social database will feature standard electronic mail, setting a profile photo, uploading images, posting content as well as following other followers. This nature of social media database includes the attributes of advertisement identity, campaign identity, Facebook campaign identity, age, gender, interest1, interest2, interest3, impressions, click, spent, total conversions and approved conversions. The social media database attributes along with the minimum range as well as the maximum range details are described in the Table 1.

SNO	ATTRIBUTES	MINIMUM RANGE	MAXIMUM RANGE
1	ad_id	708746	711877
2	campaign_id	916	916
3	fb_campaign_id	103916	104438
4	age	34	48

5	gender	Female	Male
6	Interest1	7	65
7	Interest2	8	70
8	Interest3	8	68
9	impressions	292	57665
10	clicks	0	14
11	spent	0	18.07
12	Total_conversion	1	2
13	Approved_conversion	0	1

Table 1 Social Media Database Details

For this Canopy clustering process the mixed nature of database is utilized like Simple Facebook Ad Campaign Dataset, which is attained from the kaggle database website. The sample of 380 data sets are considered and taken it as a clustering task. For this data set the total conversion number is employed with the total number of installs or signups for example while approved conversions enlighten how many signups became actual active users and how many comes under non active users. This kind of data will use to attract the consumer for projecting the business Ads. Through this way the concern can enlighten their business.

This nature of data comes under the technique of categorization. With the assistance of canopy clustering the conversion data set is categorized into active users or non active users as well as interested user or not interested user and also categorized into the age group of the consumers. These social media data details of the specific organization will lead to improve and enhance their current status of their business.

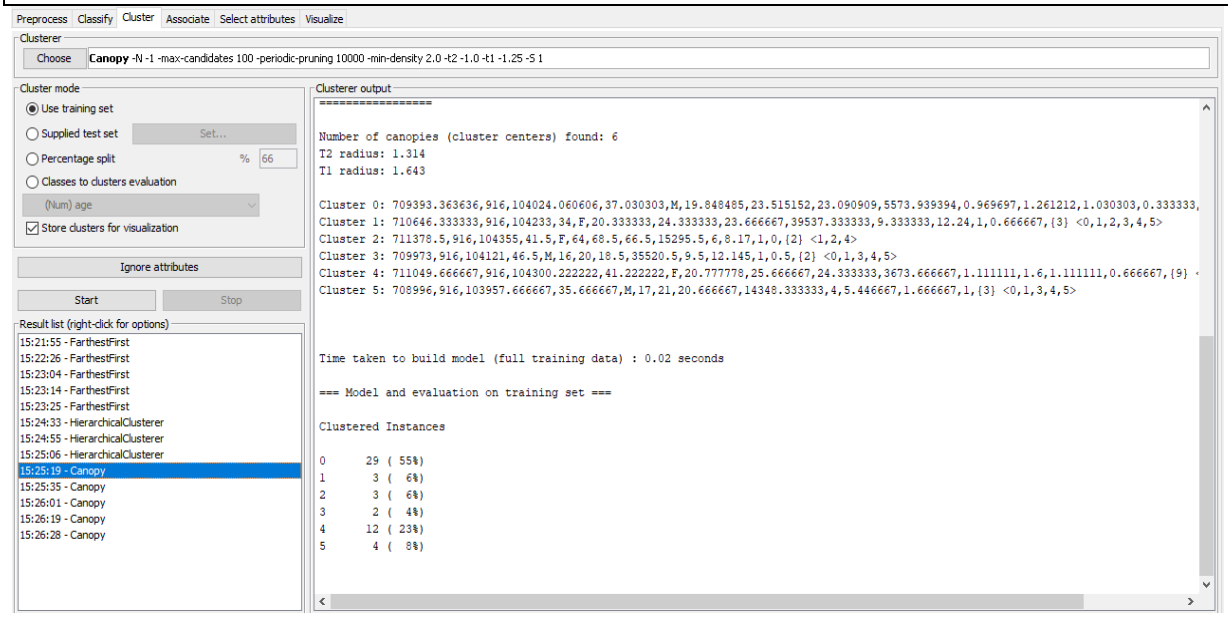


Figure 5 Sample outcomes of Social Media with six clusters

The above said figure demonstrates about the sample outcomes of the social media datasets with various numbers of clusters using the technique of canopy clustering algorithm with Weka software. The dataset of social media conversion rate is applied in the efficient clustering technique of entropy with the 13 kinds of attributes along with the count of three hundred and eighty datasets. In figure 5, number of clusters are predetermined as 6. Hence six number of canopy centroid points are to be determined for generating the clusters. Here, the threshold values T1 and T2 are determined as 1.314 and 1.643 respectively. According to the T1 and T2 values six numbers of clusters are generated. The following table discuss about the cluster instances with details of number of clusters generated, number of data points are considered for clustering as well as the percentage of specific cluster is to be expressed.

CLUSTER NO	NUMBER OF DATA POINTS CONSIDERED	PERCENTAGE
0	29	55%
1	3	6%
2	3	6%
3	2	4%
4	12	23%
5	4	8%

Table 2 Clustering details

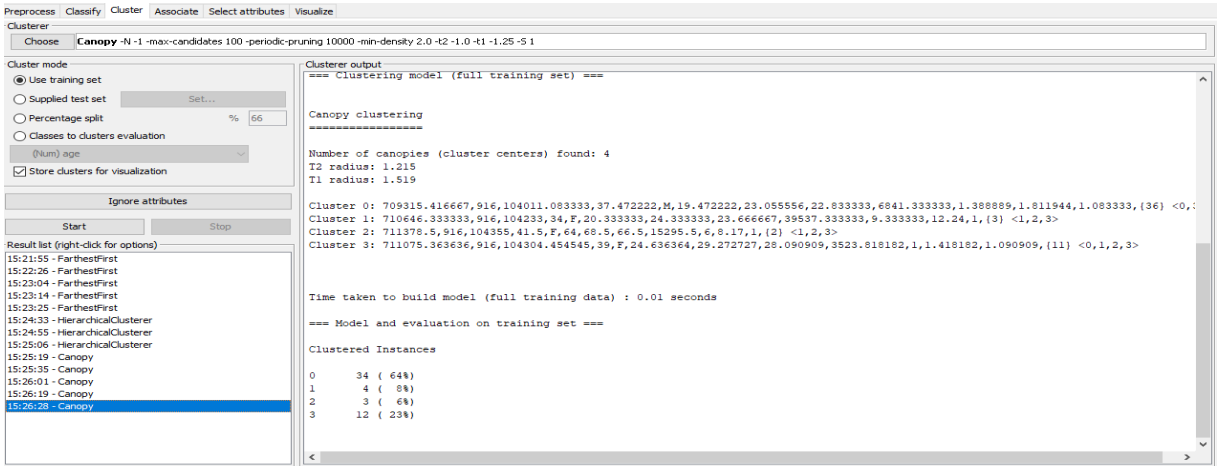


Figure 6 Sample outcomes of Social Media with four clusters

In figure 6, number of clusters are predetermined as 4. Hence four number of canopy centroid points are to be determined for generating the clusters. Here, the threshold values T1 and T2 are determined as 1.215 and 1.519 respectively. According to the T1 and T2 values four numbers of clusters are generated. The following table discuss about the cluster instances with details of number of clusters generated, number of data points are considered for clustering as well as the percentage of specific cluster is to be expressed.

Table 3

CLUSTER NO	NUMBER OF DATA POINTS CONSIDERED	PERCENTAGE
0	34	64%
1	4	8%
2	3	6%
3	12	23%

Table 3 Clustering details

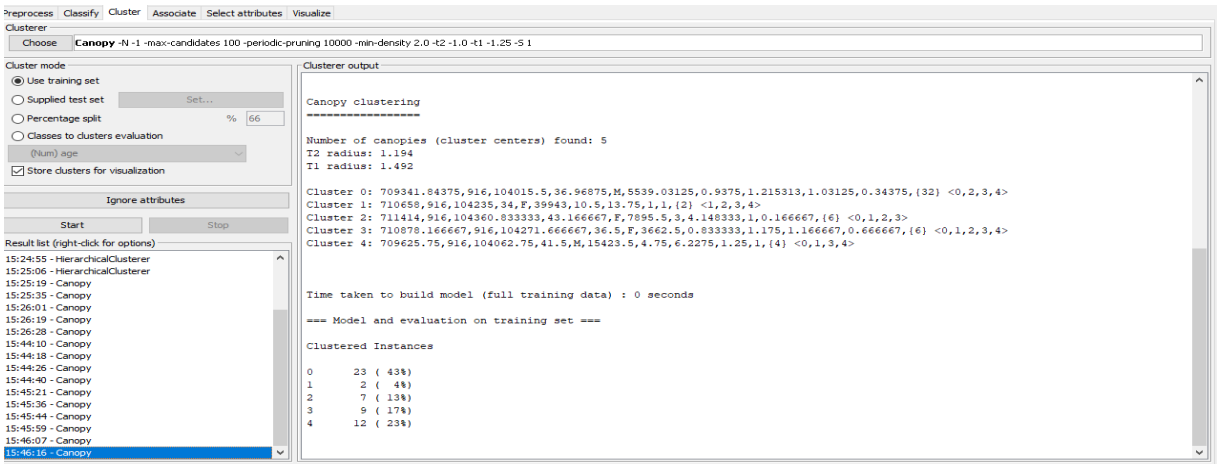


Figure 7 Sample outcomes of Social Media with five clusters

In figure 6, number of clusters are predetermined as 5. Hence five number of canopy centroid points are to be determined for generating the clusters. Here, the threshold values T1 and T2 are determined as 1.194 and 1.492 respectively. According to the T1 and T2 values five numbers of clusters are generated. The following table discuss about the cluster instances with details of number of clusters generated, number of data points are considered for clustering as well as the percentage of specific cluster is to be expressed.

CLUSTER NO	NUMBER OF DATA POINTS CONSIDERED	PERCENTAGE
0	23	43%
1	2	4%
2	7	13%
3	9	17%
4	12	23%

Table 4 Clustering details

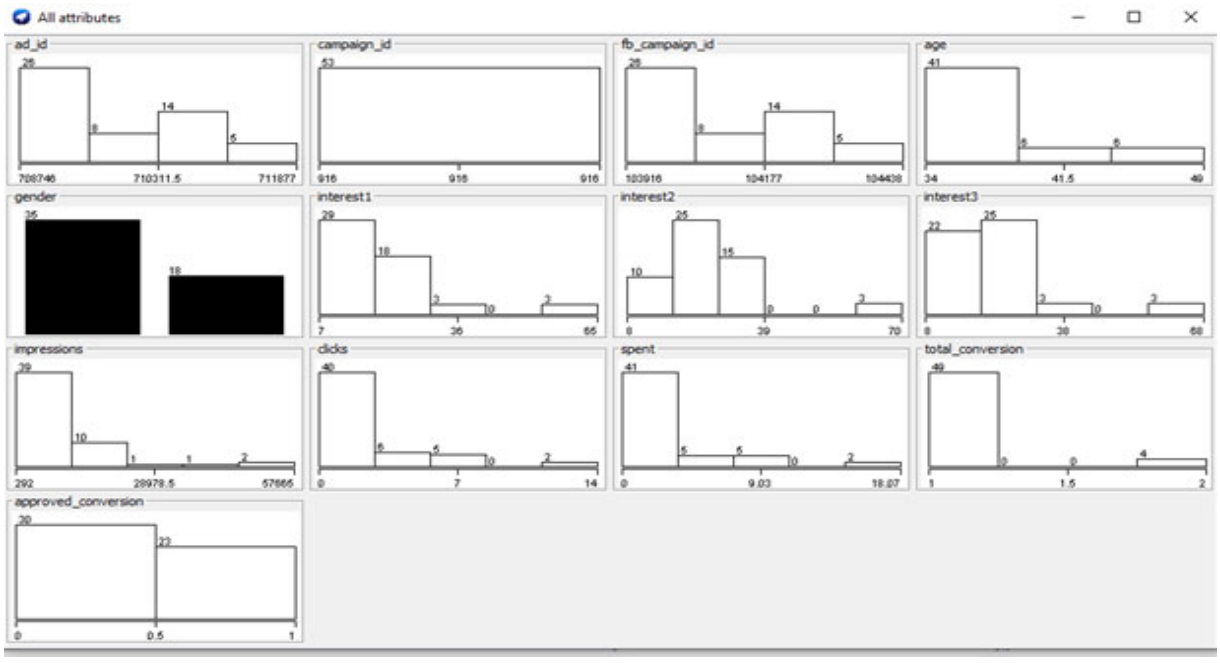


Figure 8 Sample Bar charts of social media dataset with the entire attributes

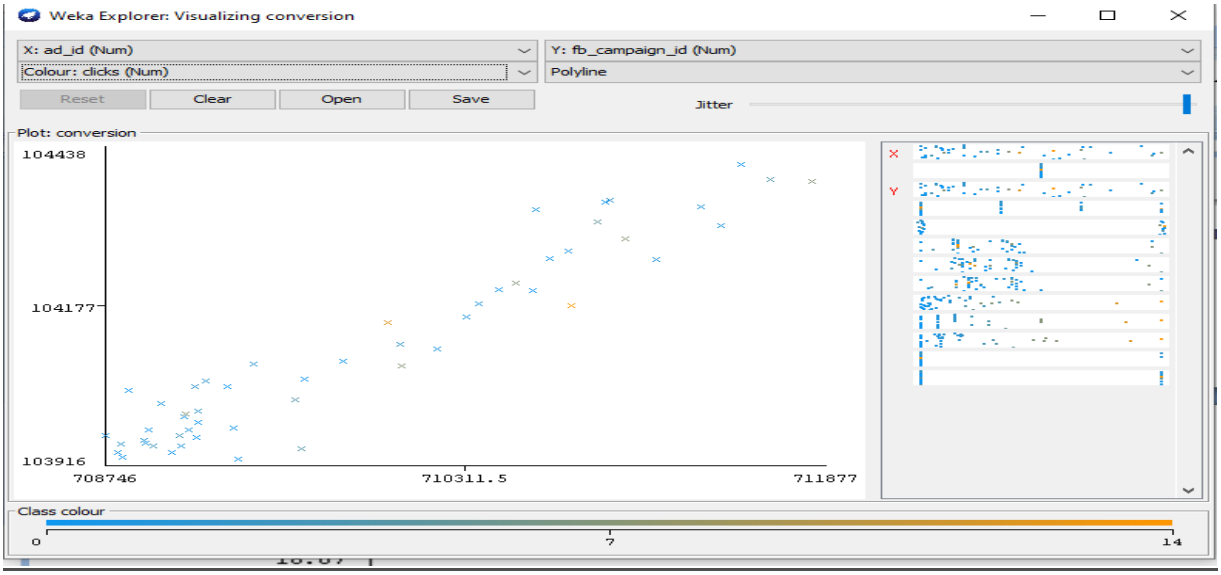


Figure 9 (a) Visualization result of social media dataset

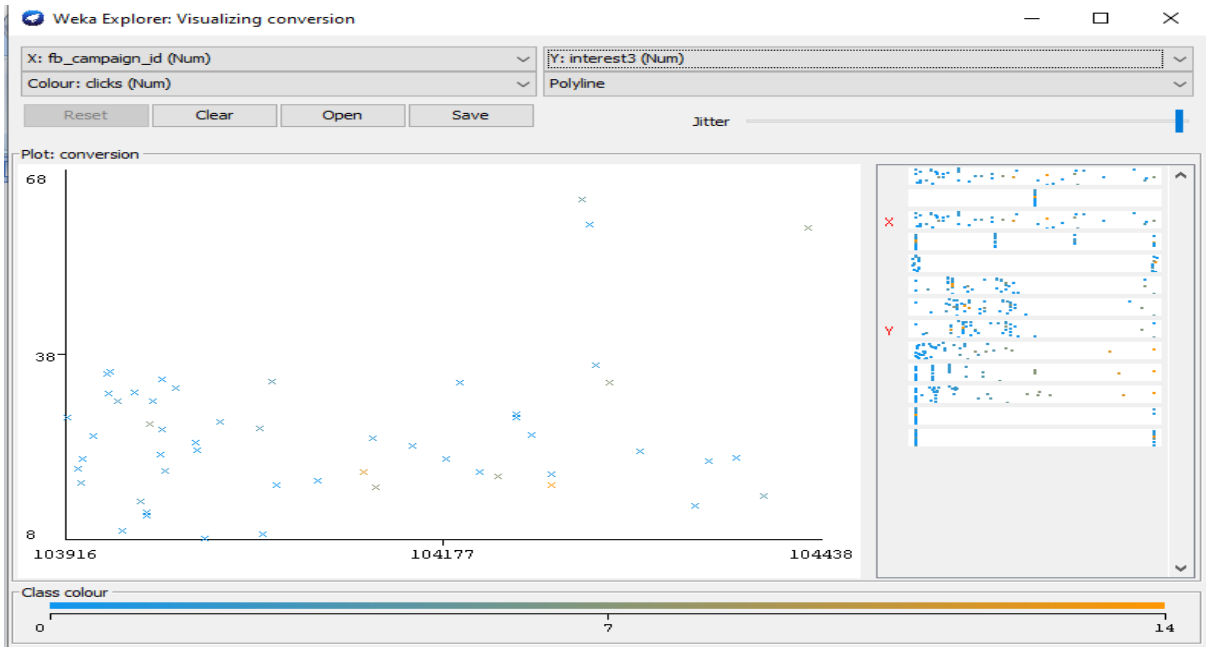


Figure 9 (b) Visualization result of social media dataset

Figure 9 demonstrates about visualization conversion result of the social media dataset with various attribute combinations using the Weka software. The first figure displays the results about the attributes of advertisement identity as x axis along with maximum and minimum range and face book compaign identity as y axis along with the values of its maximum and minimum range with the assistance of canopy clustering technique [8].

The second figure demonstrates and considers the attribute of face book compaign identity as x axis with the range of maximum and minimum as well as the attribute of interest3 as y axis with its maximum and minimum range values. The clusters framed with the assistance of canopy clustering technique.

VI. CONCLUSION

To identify the enhanced opportunities from targeted disclose advertising, to exhibit a banner advertisement to the online consumers who is majority anticipated to acquire a favourite action similar to sign up a newsletter or purchasing a product [9]. Determining the most outstanding and admirable commercial impression, it refers the opportunity to show an specific advertisement to the need of the customer needs the ability to measure the prospect that the customer who perceives the advertisement on the clients browser will obtain an achievement, that will identify the customer behaviour, mentality, taste of the product, future projections along with business improvement techniques [10]. To improve the business activities of any concern need to analyse the consumer behaviour based on the social media database analysis. With the assistance of big data clustering techniques will lead to do the real time social media dataset analysis. The most efficient canopy clustering technique is employed to do the grouping activates perfectly. The result analysis will exhibit the visualization effect of grouping process with the excellent support of the various social media attributes such as advertisement identity, Facebook advertisement identity etc, [11].Moreover with the assistance of conversion rate is gathered from the advertising of social media websites and web pages. Those details are analysed and estimated with the technique of clustering such as grouping particular age group of mankind preferences as well as their behaviour. This will lead to identify the particular consumer of the products which will enhance the productivity of any concern successfully.

In future direction of this work can be extended with measuring the performance of clustering is concentrated like estimating the accuracy of clustering performance, measuring the quality of clusters and time consumptions, to enhance the accuracy as well as comparing the accuracy with other clustering techniques.

References

[1] L. U. Khan, I. Yaqoob, N. H. Tran, S. M. A. Kazmi, T. N. Dang, C. S. Hong, Edge-computing-enabled smart cities: A comprehensive survey, IEEE Internet of Things Journal 7 (10) (2020) 10200–10232.

[2] L. U. Khan, I. Yaqoob, M. Imran, Z. Han, C. S. Hong, 6G wireless systems: A vision, architectural elements, and future directions, IEEE Access 8 (2020) 147029–147044.

[3] K. Biswas, V. Muthukkumarasamy, Securing smart cities using blockchain technology, in: IEEE 18th International Conference on High Performance Computing and Communications, Syd- ney, Australia, 2016, pp. 1392–1393.

[1] Sunjana, Aplikasi"Mining Data Decision Tree," Semin. Nas. Apl. Tknologi Inf., vol. 2010, no. Snati, pp. 1–6, 2010.

[2]L.R.AnggaGinanjarMabrur,Penerapan,"DataMiningUntukMemprediksiKriteriaNasabahKredit," J. Komput. dan Inform., vol. 1, no. 1, pp. 53–57, 2012.

[3] Amresh Kumar, Yashwant S. Ingle "Canopy Clustering: A Review on Pre-Clustering Approach to K-Means Clustering," International Journal of Innovations & Advance me nt in Computer Science, SSN 2347 – 8616 Volume 3, Issue 5 July 2014.

[4] Gongjian Zhou, "Improved Optimization of Canopy – Kmeans Clustering Algorithm Based on Hadoop Platform," ICITEE '18: Proceedings of the International Conference on Information Technology and Electrical Engineering 2018, December 2018, Article No.:19 Pages 1–6, https://doi.org/10.1145/3148453.3306258.

[5] Dweepna Garg, Khushboo Trivedi, B.B.Panchal, "A Comparative study of Clustering Algorithms using MapReduce in Hadoop," International Journal of Engineering Research & Technology (IJERT), Vol. 2, Issue 10, October.

[6] Donliang Xia, Feifei Ning, Weina He, "Research on Parallel Adaptive Canopy-K-Means Clustering Algorithm for Big Data Mining Based on Cloud Platform," Journal of Grid Computing , volume 18, pages263–273 (2020).

[7] KuldeepSingh,HarishKumar,Shakya, BhaskarBiswas, "Clustering of people in social network based on textual similarity," Perspectives in Science, Volume 8, September 2016, Pages 570-573.

[8] Dr. Punidha R, Anitha A, Arulanandan S, Karthikeyan M, "Analysis of Dataset in Social Media Using K-Means Genetic Algorithm (March 2018)," International Journal of Pure and Applied Mathematics, Volume 119, No. 15, 2018.

[9] Ahmed Alsayat, Hoda el-sayed, "Social media analysis using optimized K-Means clustering," Conference: 2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA), June 2016.

[10] Priyanka Sinha, Lipika Dey, pabitra Mitra, Dilys Thomas, "A Hierarchical Clustering Algorithm for Characterizing Social Media Users," WWW '20: Companion Proceedings of the Web Conference 2020, April 2020, Pages 353–362, "https://doi.org/10.1145/3366424.3383296"

[11] Pascal Held, Kai Dannies, "Clustering on Dynamic Social Network Data," Synergies of Soft Computing and Statistics for Intelligent Data Analysis pp 563-571.

[12] Simon Elias Bibri, "On the Sustainability of Smart and Smarter cities in the era of Big Data :an interdisciplinary and transdisciplinary Literature Review, Journal of Big Data, Springer Open Access, (2019) 6:25 https://doi.org/10.1186/s40537-019-0182-7 Volume 6, March 2019,

[13] Newman Enyioko, Gabriel A Okwandu, "Effect of Social Media Marketing on the Conversion Rate of Deposit Money Banks in Nigeria", JEMA Journal Ilmiah

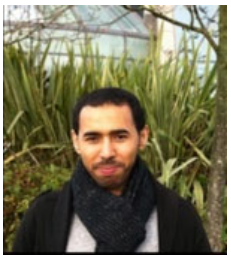
Bidang Akuntansi dan Manajemen, DOI: 10.31106/jema.v16i1.2141, March 2019.

[14] Shuai Yang, Shan Lin, Jeffrey R. Carlson, William T Ross, "Brand Engagement on Social Media will Firms' Social Media efforts in influence Search Engine Advertising Effectiveness?", Journal of Marketing Management, DOI: 10.1080/0267257X.2016.1143863, Feb 2016.

[15] Hussain Saleem, M Khawaja Shaiq Uddin, Syed Habib-ur-Rehman, Ali Muhammad Aslam,"Strategic Data Driven Approach to Improve Conversion Rates and Sales Performance of E-Commerce Websites", International Journal of Scientific and Engineering Research, April 2019.

[16] Naveen Gudigantala, Pelin Bicen, Mike Eom, "An examination of antecedents of conversion rates of e-commerce retailers", Management Research Review, Volume 39, Number 1, Jan 2016.

[17] Noor S Sagheer, Suhad A Yousif, "Canopy with k-means clustering algorithm for big data analytics", Fourth International Conference of Mathematical Sciences 2020, March 2021



Khaled H. Alyoubi is an Associate Professor of Computer Science at Faculty of Computing and Information Technology in King Abdulaziz University. His research interests include Data sciences, Data management, IR, Data Analytics and Data mining. He has a PhD in Computer Science from Birkbeck University of London, UK.

Fahd S. Alotaibi earned his PhD in Computer Science from the University of Birmingham, United Kingdom in 2015. He is currently working as an Associate Professor in the Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. His research interests include Natural Language Processing, Data Science, and Data Mining.