Clustering Based Opinion Mining for Online Shopping- COMOS

Baser Surya Kailash[1], KeravPandya[2], Zdzislaw Polkowski[3]

[1]*Gujarat Technological University, Gujarat, India*
[2]*Gujarat Technological University, Gujarat, India*
[3]*Jan Wyzykowski University, Poland*

Abstract- E-shopping is gaining a lot of popularity due to various reasons such as ease of use, availability of lot of variety, reduced shopping time etc. Thus it is very important for e-retailers to identify the opinions of the prospective customers for indulging in e-shopping. Further, it is also important to group the prospective customers on the basis of their present & past experiences of online shopping and their further experiences from the same. Opinion mining involves performing data mining tasks on the opinions of user's of some products or services. In the present work, aim is to cluster the prospective customers on the basis of their past experiences and future expectations from online shopping specifically on product delivery & online services. Partitioning based k- means clustering has been used for grouping the customers based on the opinion given by them. For this work, the opinion of customers has been collected in form of surveys done. The opinions have been categorized into three clusters which are positive, neutral, negative. Further, within each cluster data elements have been analysed on the basis of education of the respondents. To handle the visualisation of data, some dimensionality reduction technique has also been applied on it. The results show that there is a positively growing opinion in favour of online shopping over select cities of Gujarat.

Keywords- Clustering, Online Shopping, Opinion, Customer, Product

I. INTRODUCTION

Now a days, social media such as Twitter, Facebook and some of the e-commerce sites like Amazon, Flipkart, Sanpdeal, ebay allows the users to flaunt their views, opinion about the events like (Disasters, terrorist attack, sports events, etc.) and for product reviews. The number active users of social network currently exceeds 2.46 billion and is expected to reach 3.02 billion by 2021 [1]. This would nearly account for a third of the projected population. This would include nearly 750 million active users from china and around 330 million users from India. Leading social networking sites like Facebook, Twitter, etc. have high user engagement rates. The engagement rate of a user refers to the stickiness of the user or any quantifiable metric that captures the duration and frequency of how likely a particular active user would spend time on the social networking site. Currently this statistic when measured in term of duration sits at around 135 minutes per day [1].

This has resulted in the accumulation of massive amounts of diverse, unstructured data which allows the researchers to build the more robust recommendation system or can be used to gain insights regarding any specific topic, field by using this paramount information available in the social media. This task of extracting meaningful information from large pool of data is what is referred to in the literature, as Data Mining. The process of extracting the data is very complicated due to its complete dependency on the quality of data.. It includes a set of data analysing technique that not only improves the quality of data but also helps to delineate hidden relationship present in the data.

There are various types of data mining techniques that are very popular and well known. Some of them are classification, clustering, prediction etc. Broadly we classify them as supervised and unsupervised methods. Supervised methods are those methods which have predefined class labels for the data over all the set of attributes. For example classification of data according to predefined class labels is a popular supervised learning method. While unsupervised methods are the methods having no predefined class labels. For example clustering, which contains no predefined set of class labels.

There are number of application were we can use the e-reviews such as Sentiment analysis for a particular product, emotion analysis, recommendation and prediction these type of analysis may help to expand the business and to get the feedback about the products.

Sentiment analysis is nothing but is to classify the sentence in positive, negative and neutral or in other words Sentiment analysis is the task of assigning label or score to a piece of text that can be used to identify the sentiment or the opinion of the people regarding the topic being referred to in the particular sample of text. This can potentially be used to identify the opinion regarding any particular product given the reviews made of that product by its users. Thus, several businesses collect these data from customers through several outlets.

Clustering is used to cluster similar data points fall into same cluster. There are various methods of clustering such as partitioning based clustering methods, density based clustering methods, grid based clustering etc.

In this work, we have proposed an approach(COMOS) that uses partitioning based k-means clustering method to find out the opinions of customers. The objective of using clustering in this paper to minimise the intra cluster distance between them. In k -means clustering algorithm we predefined the value of k that is number of clusters. So we random the select the k data points as cluster centres. Assign all the objects to their closets data points cluster. For calculating the distance we are using the Euclidean distance function.

II. LITERATURE REVIEW

Several works had been done in the past on sentiment analysis[1] using data mining. Yang et al.[2] used fuzzy clustering method for microblogs[3]. Lei et al. [4] used sentiment dictionary and machine learning methods to classify emotions on Twitter data . Hoque et al.[5] used Support Vector Machine (SVM), Hidden Markov Models (HMM) etc. for binary classifications between different stimuli patterns such as smiles.

Agarwal et al.[6] uses the lexicon based approached for sentiment analysis over the twitter data. In this author perform the geo-spatial sentiment analysis and its finding shows most influential country for that particular event, author crawled the tweets regarding the Brexit event. Geo-spatial sentiment type of analysis is important because to know the correct source of the target users if the similar events is happen in future. In [7,16] author find the semi-automatic retrieval of frequently asked questions in Short Message Services. Due to word limit in SMS author uses the short form, abbreviations etc. Similarly in tweets users post the tweets which contain a lot of noise in the tweets. So we have the fallow the same text pre-processing steps in this work.

McDuff et al. [8] proposed a different framework to analyze facial responses in the media content. Dey et al.[9] applied Naive Bayes and K-NN Classifier for analyzing reviews of different entities and products. Lei et al.[10] also used machine learning methods to classify polarity in twitter based text. However, sentiment analysis over product review is quite a recent advancement in the field of text mining. Brenden et al.[11][12] found in research that, product reviews had a very high correlation with sentiments. Cindy et al.[13] proposed a statistical method for social network event topic mining.

Sentiment Analysis is a classification task that attempts to classify a given sample into one of several classes that indicate the sentiment expressed in the sample text [14]. Most of the work in this regard focuses on lexicon based approaches. In these approaches each word is assigned a score based on the connotation of the word in the English language. Through clever aggregation of these scores for the entire sample the sentiment score for the entire sample is generated. Based on this score the given sampled is classified into an appropriate class that indicates its sentiment. There are several ways in which the sentiment score is aggregated ranging from simple summation of scores for individual words to complex functions obtained by identifying the sentiment for a phrase and then aggregating these scores using an appropriate function.

Recommendation systems work on the principle of how likely a given product may be recommended to the user based on the previous choices made by the same or users with similar usage history. These systems try to maximise the probability of the purchase of the recommended product. Turney [15] describes a recommendation system based on the unsupervised learning approach that classifies reviews as “thumbs up”, meaning recommended to “thumbs down” meaning not recommended. This is done with the help of identifying the semantic orientation of the individual phrases

A phrase is said to have good orientation if the mutual association between the individual words that make up the phrase has a net positive score (E.g.: “subtle nuance”) or a net negative score (E.g.: “very cavalier”). Based on the similarity calculated by aggregating the score obtained for individual phrases the reviews are grouped into several categories where each categories indicate samples that can be recommended when one of the other sample is considered [15].

In [17], Valence Aware Dictionary for Sentiment Reasoning (VADER), which is a rule-based model for sentiment analysis is presented. VADER uses a lexicon list with sentiment measures to compute sentiment score for textual data. Sentiment analyzer examines the textual input to mine the feeling of user, which is present in the text and provides sentiment information indicative of feelings expressed by the users. Customer reviews for products collected through e-commerce sites, blogs, social media etc., have rich knowledge about the products and their usability. Sentiment analyser can be used to undermine sentiment information from the customer reviews. A large number of reviews being collected every day, better schemes are highly desirable to analyse sentiments of reviews and to provide visualisation of sentiment information.

In this paper, we have analysed data of online shopping over past experiences and future expectations of users based upon questionnaires. K-means clustering is used for pattern analysis and PCA have been used for visualization.

III. METHODOLOGY

In this section, we discuss the methodology of the proposed approach

3.1 Data Collection

The very first step in any data analysis process is to collect data suitable for the analysis. In this paper, we have used the dataset consisting of customer's satisfaction reviews for online shopping and online payment process. Each customer has rated the services in the range of 1 to 5 where, value 1 reflects dissatisfaction and value 5 reflects complete satisfaction. Further, additional information such as education, occupation, annual family income, mode of payment is also gathered from customers.

3.2 Data Pre-Processing

Data collected from the real word is highly likely to contain several errors such as missing values, inconsistencies [19], formatting errors[20], redundant values etc. Data pre-processing[19] aims at handling such discrepancies from data to make it suitable for task of data analysis. Data pre-processing involves various steps such as data cleaning [21] (smoothening data), data transformation

(converting data into format suitable for analysis/ data normalization), data integration (resolving data conflicts/ combining data from multiple sources), data discretization[21] (dividing data into groups) and data reduction (finding features of interest from the data/capturing max. data variance with less number of attributes).

3.3 Clustering

Data Clustering [21][22] is the method of grouping similar data points together, with similar attributes. It aims at finding natural groups in the data in such a way that objects belonging to one cluster must have high similarity as compared to objects belonging to other clusters. In other words, inter-cluster similarity should be more than intra-cluster similarity. Objects belonging to the same cluster shows some pattern which is common across that cluster. There are various methods of clustering such as partitioning based clustering method [21], density based clustering methods [21], grid based clustering methods [23] etc. In this work, we have used partitioning based clustering method K-means. K-means [21] is one of the most popular and simple clustering algorithm. The various step involved in K-means clustering algorithm are as follows:

Input: Initial set of points

Output: Set of K cluster assignments

Step 1: Randomly chose K cluster centroids, where k is predefined.

Step 2: Assign input points to their nearest cluster centre based upon Euclidean distance.

.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \qquad (1)$$

Step 3: Compute the mean of all points for each of the cluster. Make these mean values as new cluster centres.

Repeat steps 2 and 3 until there are no changes in the cluster centre assignment.

3.4 Results and Knowledge Discovery

The next step after clustering algorithms is to find out the overall rating of online shopping and payments services by different customer's. The overall goal of this step

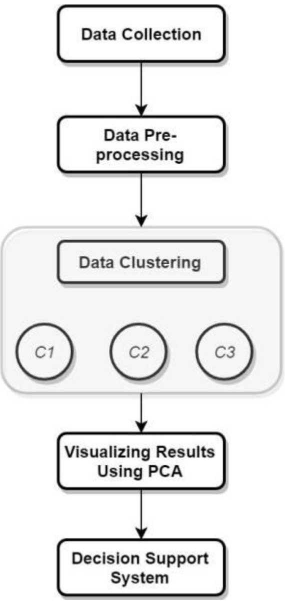

Figure 1: Flow chart of the proposed methodology

is to find out the patterns of interest from the data in such a way that it can be used to improve the services.

IV. APPLICATION OF THE PROPOSED APPROACH

The methodology discussed in the previous section is applied to the dataset collected by means of questionnaire. In addition of various personal information fields such as name, occupation, annual income and education, several other aspects such as user's past experience of online shopping, future expectations from online shopping, past experiences with online payment & delivery, future expectation with online payments & delivery were also included. The users are required to fill a value in a range of 0 to 5 where 0 corresponds to strongly disagree (Negative reviews) and 5 corresponds to strongly agree (Positive reviews) for a service.

In the present work, we have used K-means clustering algorithm to find out hidden patterns in the data. The value of k has been chosen to be three because the feelings & opinions of the users of online shopping can be categorized into three groups which are- positive, negative and neutral. Initially, we used K -means clustering algorithms to find out the changes in users opinions i.e. whether online shopping is more convenient, enable to chose from wide variety, more reliable etc. in the past to their expectations in the future. As in our case input dataset is multi-dimensional, so it is not possible to directly visualize the K-means clustering results. Therefore in order to visualize the clustering results, we have used Principal Component Analysis(PCA) [24] algorithm to find out the dimension that captures most of the variance of the data.

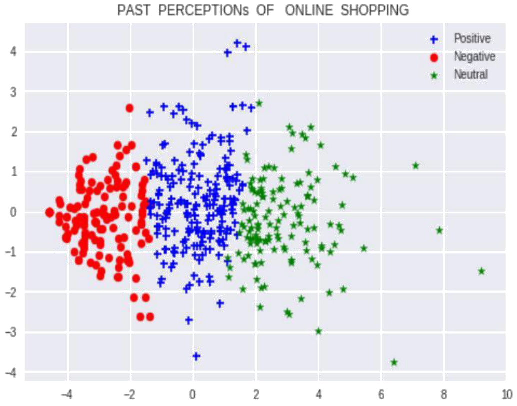

Figure 2(a): Clustering Results of Customers Past Perceptions of Online Shopping

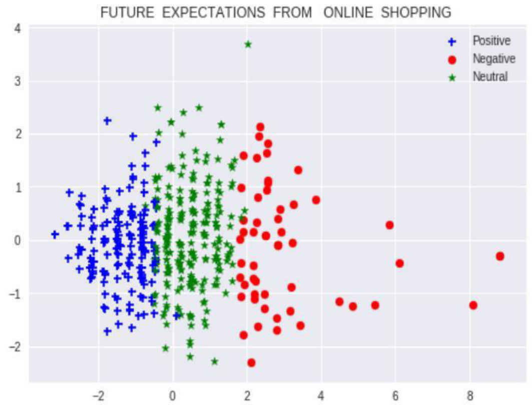

Figure 2(b): Clustering Results of Customers Future Expectations of Online Shopping

PCA[24] algorithm forms it basis on Eigen values and Eigen vectors. It starts out by finding a set of orthogonal vectors in the data. Each of the vector have associated Eigen values that helps in determining the importance of that vector in capturing variations in data features. Higher Eigen value corresponds to the most important dimension. So we use only those vectors that are able to explain most of the data variations. In the present work, we have used two PCA components to visualize the results of K-means clustering results. Figure 2(a) & 2(b) shows the results of the K-means clustering algorithm. Further, we used K-means clustering algorithm to find out users opinion/review on online payments & delivery of products i.e. whether credit card details are safe, home delivery is safe, whether product received is same as that of ordered etc.

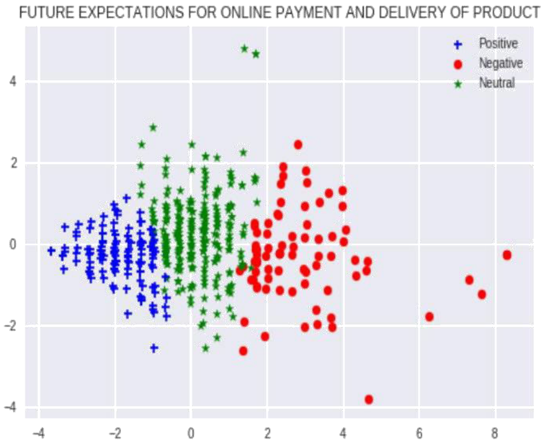

Figure 3(a): Clustering Results of Customers Past Perceptions of Product Delivery & Online Payments

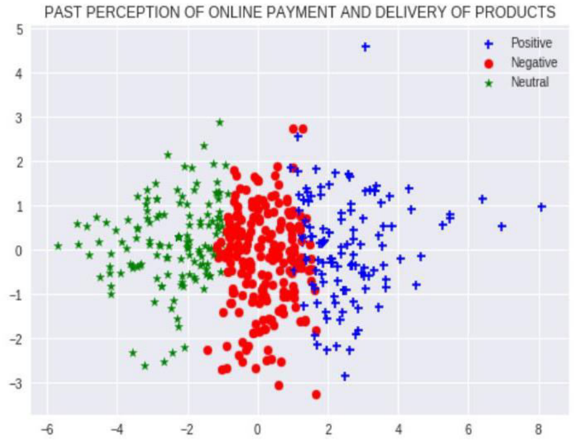

Figure 3(b): Clustering Results of Customers Future Expectations of Product Delivery & Online Payments

Furthermore in this work, we have also explored the opinions of UG and PG students for online shopping and payment system with the help of education qualification attributes. Figure 5(a) shows the opinion of students for the past online shopping services such as whether online shopping is more convenient, enable to chose from wide variety, more reliable etc. Figure 5(b) shows their opinions regarding future expectation for online shopping services. Figure 6(a) shows the students past opinions on the online payment and delivery services such as whether credit card details are safe, home delivery is safe, whether product received is same as that of ordered etc and Figure 6(b) shows their opinion regarding

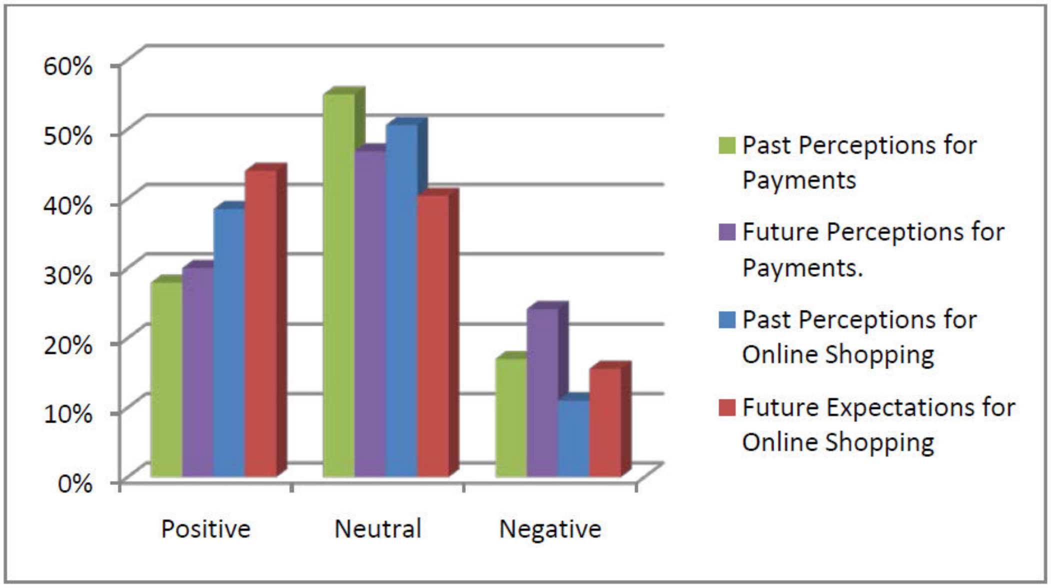

Figure 4: Percentage(%) changes in the Customers opinions regarding online shopping product delivery & online payments

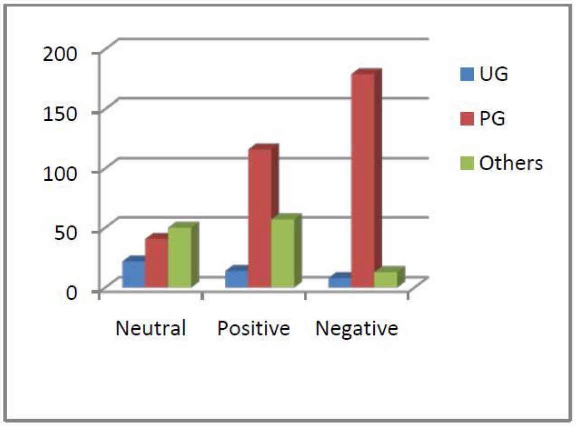

Figure 5(a):UG, PG & Other Students Past Perceptions of Online Shopping

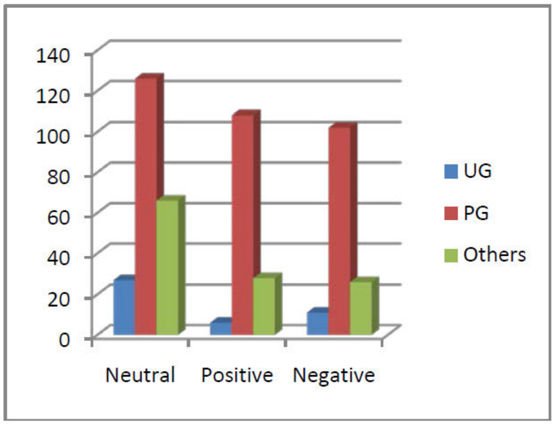

Figure 5(b): UG, PG & G Students Future Expectations of Online Shopping

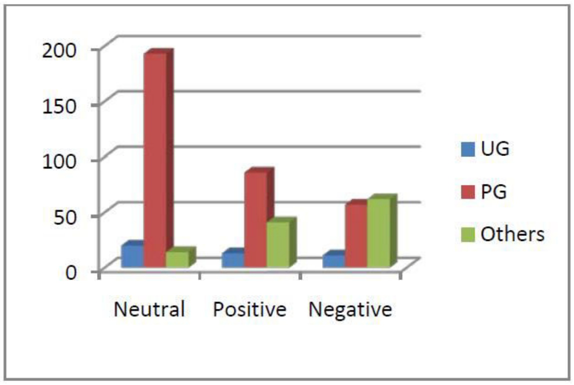

Figure 6(a): UG, PG & other students Past Perceptions of Online delivery & Payments Services

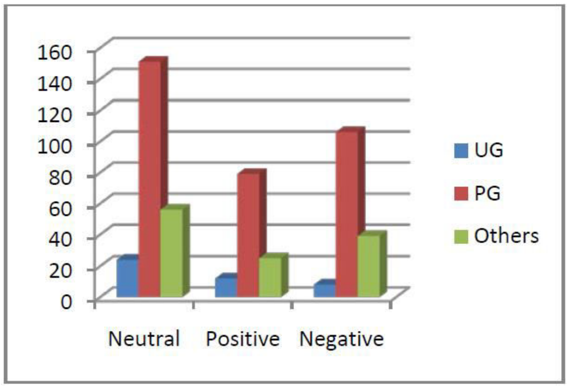

Figure 6(b): UG, PG & other students Future Expectations on Online delivery & Payments Services

TABLE I. Clustering Results of k-means for customer's opinion for online shopping

S.No	*Description*	*Cluster No.*	*Positive*	*Overall*	*Trend*
1	Past Perceptions of Online Shopping	C1: Positive C2: Neutral C3:Negative	192 253 55	Neutral	Change in overall trend from neutral to positive
2	Future Expectations of Online Shopping	C4: Positive C5: Neutral C6:Negative	220 138 142	Positive	
3	Past Perceptions of Product Delivery & Online Payments	C7: Positive C8: Neutral C9:Negative	140 275 85	Neutral	Slight increase in both, positivity and negativity
4	Future Expectations of Product Delivery & Online Payments	C10: Positive C11: Neutral C12:Negative	145 234 121	Negative	

Table II. Educational qualification wise results for customers' opinions for online shopping

Opinion Type	*UG*			*PG*			*Others*		
	Past Perception of online shopping	*Future Expectations of online shopping*	*Overall Trend*	*Past Perception of online shopping*	*Future Expectations of online shopping*	*Overall Trend*	*Past Perception of online shopping*	*Future Expectations of online shopping*	*Overall Trend*
Positive	*14*	*6*	↓	*116*	*118*	↑	*57*	*28*	↓
Neutral	*22*	*27*	↑	*41*	*116*	↑	*50*	*66*	↑
Negative	*8*	*11*	↑	*179*	*102*	↓	*13*	*26*	↑

Table III. Educational qualification wise results for customers' opinions for Product Delivery & Online Payments

Opinion Type	*UG*			*PG*			*Others*			
	Past Perception of product delivery and online payment	*Future Expectations of product delivery and online payment*	*Overall Trend*	*Past Perception of product delivery and online payment*	*Future Expectations of product delivery and online payment*	*Overall Trend*	*Past Perception of product delivery and online payment*	*Future Expectations of product delivery and online payment*	*Overall Trend*	
Positive	13	12	↓	86	89	↑	41	25	↓	▼
Neutral	20	24	↑	193	151	↓	62	56	↓	▼
Negative	11	8	↓	57	96	↑	17	39	↑	↑

future expectation from these services. Table II. & III shows the results of undergraduate, postgraduate and others for online shopping, product delivery and payment services based upon their opinions.

V. CONCLUSION

Opinion mining is a subarea in the field of data mining. In the present work, the focus is to perform opinion mining based on the data collected in the form of surveys from prospective customers for online shopping from select cities in Gujarat. K-means clustering is very popular clustering technique which finds a lot of applications including grouping of survey data for online shopping. The value of k has been chosen to be three because the feelings & opinions of the users of online shopping can be categorized into three groups which are-positive, negative and neutral. After the clusters have been formed, further analysis have been done to identify the educational qualification of the respondents versus opinions. Analysis has also been done on the basis of the financial income of the respondents. It can be concluded that generally the customers in the state of Gujarat have a positive opinions about online shopping services. However, the respondents in Gujarat have generally shown a negative opinions towards their experience in the usage of online payments.

REFERENCES

[1] Cambria, Erik, et al. "SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives."Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016.

[2] Yang, Li, Xinyu Geng, and Haode Liao. "A web sentiment analysis method on fuzzy clustering for mobile social media users." EURASIP Journal on Wireless Communications and Networking 2016.1 (2016): 128.

[3] X Song, 5W features analysis in micro-blog. J CHIFENG Univ. 29 (1), 96:98 (2013).

[4] L Zhang et al., Combining Lexicon-Based and Learning-based Methods for Twitter Sentiment Analysis. HP Laboratories Technical Report, 2011

[5] M E Hoque, D J McDuff, R W Picard, Exploring temporal patterns in classifying frustrated and delighted smiles. IEEE Trans Affect Comput. 3(3),323-334 (2012).

[6] Agarwal, A., Singh, R. and Toshniwal, D., 2018. Geospatial sentiment analysis using twitter data for UK-EU referendum.Journal of Information and Optimization Sciences, 39(1), pp.303-317.

[7] Agarwal, A., Gupta, B., Bhatt, G. and Mittal, A., 2015, December. Construction of a Semi-Automated model for FAQ Retrieval via Short Message Service. In Proceedings of the 7th Forum for Information Retrieval Evaluation (pp. 35-38). ACM.

[8] DJ McDuff, R el Kaliouby, RW Picard, Crowdsourcing facial responses to online videos. IEEE Trans. Affect. Comput.3(4), 456–468 (2012)

[9] Dey, Lopamudra, et al. "Sentiment Analysis of Review Datasets Using Naive Bayes and K-NN Classifier." arXiv preprint arXiv:1610.09982 (2016).

[10] A Go, R Bhayani, L Huang, Twitter sentiment classification using distant supervision. Cs224n Project Report, 2009, pp. 1–12

[11] B O'Connor, R Balasubramanyan, BR Routledge, NA Smith, From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series, Conference: Proceedings of the Fourth International Conference on Weblogs and Social Media, 2010, pp. 122–129

[12] Dabral, S., Agarwal, A., Kumar, S. and Gautam, B., Passenger Abnormal Behaviour Detection using Machine Learning Approach.

[13] CX Lin et al., PET: A Statistical Model for Popular Events Tracking in Social Communities. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, pp. 929–938

[14] Fang and Zhan Journal of Big Data – "Sentiment analysis using product review data"- Journal of Big Data, springer 2015 2,5

[15] Peter D. Turney -"Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews" Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 417-424

[16] Goyal, P., Kukreja, T., Agarwal, A. and Khanna, N., 2015, March. Narrowing awareness gap by using e-learning tools for counselling university entrants. In Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances in (pp. 847-851). IEEE.

[17] C.J. Hutto and Eric Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text,"Eighth International AAAI Conference on Weblogs and Social Media, 2014.

[18] A new ANEW: Evaluation of a word list for sentiment analysis in microblogs", Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages 718 in CEUR Workshop Proceedings.

[19] Famili, A., Wei-Min Shen, Richard Weber, and Evangelos Simoudis. "Data preprocessing and intelligent data analysis." Intelligent data analysis 1, no. 1 (1997): 3-23.

[20] Hand, David J. "Principles of data mining." Drug safety 30, no. 7 (2007): 621-622.

[21] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

[22] Jain, Anil K. "Data clustering: 50 years beyond K-means." Pattern recognition letters 31, no. 8 (2010): 651-666.

[23] ZHAO, Hui, Xi-yu LIU, and Hai-qing CUI. "Grid-Based clustering algorithm." Computer Technology and Development9 (2010): 021.

[24] Maćkiewicz, Andrzej, and Waldemar Ratajczak. "Principal components analysis (PCA)." Computers & Geosciences 19, no. 3 (1993): 303-342.