Analysis and Exploitation of Twitter Data Using Machine Learning Techniques

Ganeshayya Shidaganti, Rameshwari Gopal Hulkund and S. Prakash

Abstract In the present era, Internet is a well-developed technology that supports most of the social media analysis for various businesses such as marketing of a product, analysis of opinions of different customers, and advertising most of the brands. This gathered huge popularity among different users with a fresh way of interaction and sharing the thoughts about the things and materials. Hence, social media comprises of huge data that categorizes the attributes of Big Data, namely volume, velocity, and variety. This leads to the research work of this huge data related to different organizations and enterprise firms. To analyze the demands, customer's efficient data mining techniques are required. Nowadays, twitter is the one among the social networks which is dealing with millions of people posting millions of tweets. This paper exemplifies the data mining with machine learning techniques such as TF-TDF and clustering algorithms such as hierarchical clustering, *k*-means clustering, *k*-medoid clustering, and consensus clustering along with their efficiencies.

Keywords Twitter data · Machine learning technique · Consensus clustering Big data · Social media · TF-IDF · *K*-medoid clustering

G. Shidaganti (✉) · R.G. Hulkund
Department of Computer Science & Engineering, M.S. Ramaiah Institute of Technology, Bangalore 560054, India
e-mail: ganeshayyashidaganti@msrit.edu

R.G. Hulkund
e-mail: h.rameshwari.109@gmail.com

S. Prakash
Department of Computer Science & Engineering, Dr. Ambedkar Institute of Technology, Bangalore 560056, India
e-mail: prakash.hospet@gmail.com

M.S. Reddy et al. (eds.), *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications*, Advances in Intelligent Systems and Computing 628, https://doi.org/10.1007/978-981-10-5272-9_13

1 Introduction

Social medias [1] such as twitter, Facebook, MySpace, LinkedIn, and many more are being popular in this era of Internet of everything. These microblogging sites are very advantageous to business firms, service providers, and customers. The service providers give the opportunity to enterprises to advertise their applications and products to customers through these sites. The interested customers can easily get the information about these things. This saves their valuable time. Even multiple jobs and requirements can also be posted in these sites. These social media has attracted data scientists to study the relational, social, and behavioral aspects between social sites and their implications on society. Social network analysis [2] provides opportunity to understand the interaction between individuals and group of people and communities of different networks.

Twitter is the microblogging site which is acquiring more and more popularity and growing faster. The users post their messages as tweets, and hence, per day millions of tweets are being posted. Users use this site to update what is there in their mind as status and discuss about the products and services with their relatives and friends who are staying far. This is an example for real-world scenarios like reviewing about the electronic goods, clothing, automobiles, movies to be watched, hotels, and restaurants. This site has become very useful to marketers as they will easily get to know about the customers' satisfaction related to their products and services.

Twitter sentiment analysis [3] is done to analyze the sentiments of users. Sentiments mean thoughts in positive, negative, or neutral forms. The emotion-rich data are gathered from twitter. This work includes the analysis of effectiveness of machine learning techniques on twitter corpus. This dataset is continuous. Dataset on movie reviews is discrete one. It is simple to implement machine leaning techniques on discrete datasets compared to continuous datasets. Due to limited number of characters posting, people end up with short form of words and use emoticons which give different perspective for word context.

In this paper, we have considered different clustering algorithms and other machine learning techniques. The organization of paper is as follows: Sect. 2 gives information about data preprocessing, and Sect. 3 represents experimental results.

2 Data Preprocessing

Since analysis of microblogging sites has got more importance during crises, the analysis of data is very important. Hence, data preprocessing [4] is very necessary in data mining. Therefore, the phrase "garbage in and garbage out" is specially meant for machine learning and data mining processes. During the collection of huge data, data get jumbled in different impossible forms which give informal

meaning. These kinds of things are needed to be clarified to produce meaningful and tactic results from the corpus. It also improves the quality of the data.

The prediction of knowledge during initial phases of data training becomes difficult when redundant and irrelevant data or noise is present as a part of collected data. In this paper, the noisy data are referred to URLs and stop words of English literature. This leads to maximal wastage of time during data preparation and filtering. Data preparation phase is the second phase of Big Data life cycle which includes cleaning, normalization, transformation, feature extraction, and selection of data processing techniques. The final set of the process is the processed data ready for further actions without any inconsistency. The preprocessing follows:

A. Special characters removal: The emoticons in the text file appear like the set of special characters. In certain applications or tasks, emoticons are not needed, and hence, these characters are removed from the datasets.
B. Identifying uppercases: Slang words such as BTW which is meant as "by the way," tomorrow as 2MRW, LOL, ROFL have to be either replaced or removed forever from the datasets.
C. Alphabet lower casing: In the twitter [5] dataset, most of the words are written in capital letters to highlight those words. For example, instead of writing "hello," it could be represented as "HELLO." Therefore, it is very important part of the data preparation phase of life cycle. Before removing the cases, capital letters are identified. In microblogging, even irregular casing exists as "TwInLkIIngofSTARS."
D. Compression of word: Sometimes, few words are simply exaggerated. For an instance, happy is exaggerated as "hhhaaappyyyyyy." This word contains irrelevant letters which are absolutely not needed. To increase the accuracy, the identification of the sense of a sentence is essential.
E. Identifying pointers: Pointers refer to usernames and Hash tags. In twitter, character "@" is being used to point out a particular person in their posts. To differentiate words, "#" is used instead of white spaces like "#Happy#Journey#Ishan."
F. Synset [6] detection: Synset finding is done on the words such as "create," "creation," "created," "creating" which are relevant to the word "create." Therefore, these words when appear are considered as the word "create," which is the base word. This reduces the feature vector size while preserving the worthy key terms.
G. Link removal: The URLs that are downloaded as a part of dataset are not useful for sentiment analysis and applying machine learning [7] techniques. These do not contribute anything to data mining. Hence, links are considered as garbage.
H. Stop word removal: In any natural language processing tool, the most important task is identifying the stop words and removing them.
I. Spell checking [8]: Usage of acronyms has become the trend, and in most of the microblogging sites, the number of characters is limited to 140 words. Hence, the shortened words have to be modified to the original words with the help of English dictionary.

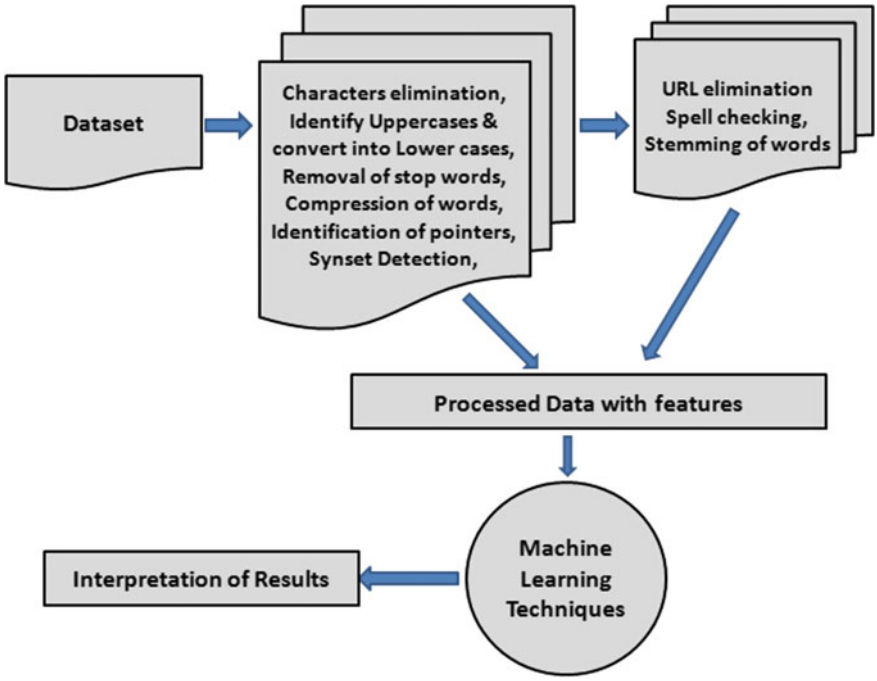

Fig. 1 Data preparation phase

J. Stemming of words: The term stemming is used in identifying the morphology of structure of given language's morphemes. It is also used to do information retrieval for the infected words to their word stems. In *R* programming, tm-package is used for stemming of words. For example, "engineer" is the root or stem word for "engineered," "engineering," and "engineers."

The data preprocessing steps are clearly shown in Fig. 1.

3 Machine Learning Techniques

As twitter data are unstructured data, to make it structured and apply some rules for further processing, machine learning comes into picture. Data refer to recorded data, whereas information refers to patterns underlying the data. To obtain the structural description of the data, the following techniques are used.

i. **TF-IDF**: Its elongated form is term frequency-inverse document frequency. This is used to check the number of times a word is repeated in a set of data. Based on the frequency of the word occurred in the different groups, categorization is done for an article. TF refers to how many times the word has occurred in an article. The term frequency for a word in an article means the ratio of the word count to the total number of words in the article. IDF describes the existence of a word in different documents as a common word between the documents. It is helpful in analyzing the different documents or article based on a single or multiple common words. For this paper, it is very helpful to analyze the tweets which share the same information based on the IDF terms.

ii. **Clustering**: Clustering is the technique used for statistical data analysis. It is needed in grouping the elements which are more similar to each other compared to other groups. In this paper, two clustering techniques are used.

A. **Hierarchical clustering**: To build the hierarchy of a statistical data, hierarchical clustering is used. The strategies for this are as follows:

- Agglomerative: It follows "bottom up" approach. Each element starts from its own cluster and pairs with other clusters which share near characters.
- Divisive: It follows "top down" approach. In this type, each element starts in one cluster are splits up as it moves further down the line.

In most of the information retrieval projects, agglomerative algorithms [4] are used rather than divisive algorithm. To do split and merge, greedy algorithm is applied. This paper work has used agglomerative algorithm. Metrics refers to the measurement of distance between the points. Some of the metrics are listed in Table 1.

In this paper, we are dealing with the Manhattan distance between the points.

A. ***k*-medoid clustering**: It is partition-based clustering algorithm. This clustering algorithm aims to distribute n observations into k clusters, in which each element belongs to the cluster of nearest mean. Euclidean distance is used as the metric. It uses PAM (Partitioning around Medoid) algorithm. PAM is faster than the exhaustive search because it uses greedy search. This algorithm follows the following procedure:

Table 1 Metrics for hierarchical clustering

Names	Formula
Euclidean distance	$\|\|a-b\|\| := \sqrt{\sum_i [\![(a\_i - b\_i)]\!]^{\wedge 2}}$
Manhattan distance	$\|\|a-b\|\| := \sum_i \|a_i - b_{i\|}$
Maximum distance	$\|a-b\| =:= \max\|a_i - b_i\|$

1. Initialize or randomly select the value of *k*.
2. Find association of each data point to its nearest neighboring clusters.
3. While the cost of configurations of data decreases,
4. For each medoid (*m*) data point and non-medoid data point (*o*),

   (i) Swap *m* and *o* and recalculate the cost.
   (ii) In the previous step, if the cost increases, then undo the swap that has happened in the previous step.

B. **Consensus clustering**: This clustering type is recommended for huge datasets. The main advantage of this approach is that these provide a final partition of data that is comparable to the best existing approaches, yet scale to extremely large datasets. Consensus clustering combines the advantages of many clustering algorithms.

4 Implantation with Results

Now, the preprocessed data can be used to interpret the results. Word cloud is used to show the importance of the words. The word cloud is formed from the term-document matrix. The representation of word cloud is shown in Fig. 2.

Fig. 2 Word cloud

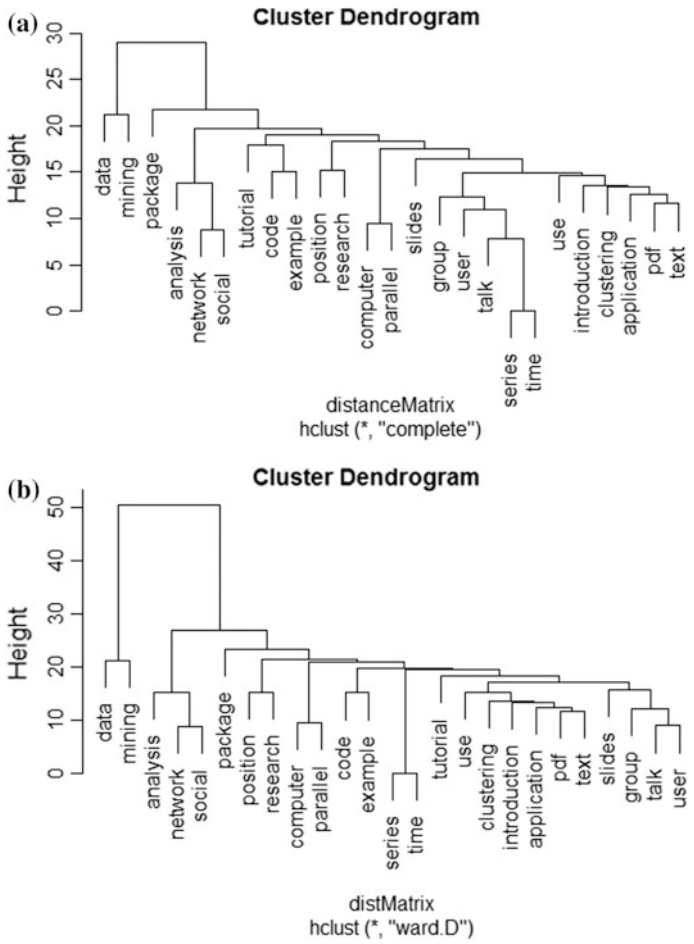

Fig. 3 Ordering of objects by hierarchical clustering, **a** with method as complete, **b** with method as "ward.D"

The analysis of hierarchical clustering shows that smaller clusters are generated. It also arranges the objects in certain orders. This is illustrated in Fig. 3.

TF-IDF [5] result gives the plot of number of counts versus terms in the dataset, as shown in Fig. 4, and Fig. 4b shows the bar plot of top few words which have occurred frequently.

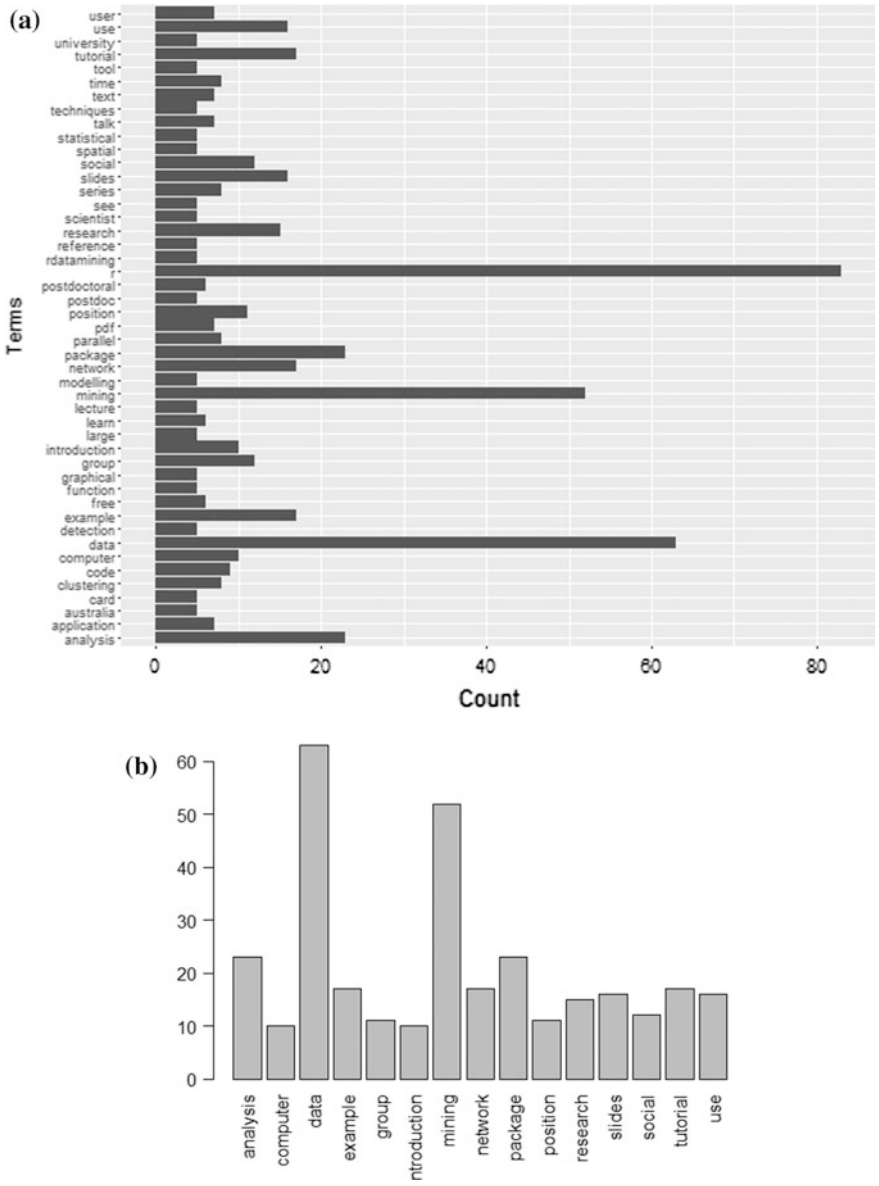

Fig. 4 **a** TF-IDF, and **b** bar plot of the terms which have occurred maximum no. of times

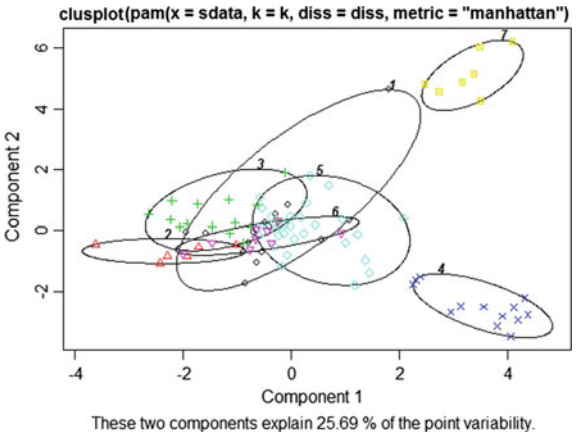

Fig. 5 *k*-medoid clustering

k-means and fuzzy [9] *c*-means(*c*-means centroid) minimize the squared error criteria and are computationally efficient. These algorithms do not require the user to specify the parameters.

There is no much difference between *k*-means and *k*-medoid clustering algorithms. *K*-medoid choose data points as centers (medoids). This algorithm is more robust to noise and outliers compared to *k*-means algorithm. It minimizes the sum of dissimilarities instead of sum of squared Euclidean distances. The common realization of *k*-medoid clustering is Partitioning around Medoid (PAM). The result for *k*-medoid algorithm is shown in Fig. 5.

The three disadvantages of these above-mentioned algorithms are as follows:

- Entities must be represented as points in *n*-dimensional Euclidean space.
- Objects have to be assigned to their respective clusters.
- Clusters must have same coordinates and must be of same shape.

To overcome these disadvantages, consensus clustering is implemented in this paper. The below graphs show the results of the same in Fig. 6.

The cumulative distribution function for the whole dataset is shown in Fig. 7.

The overall consensus clustering algorithm's cluster consensus is given in Fig. 8.

Figure 9 shows scatter plots of confidence and lift with respect to support.

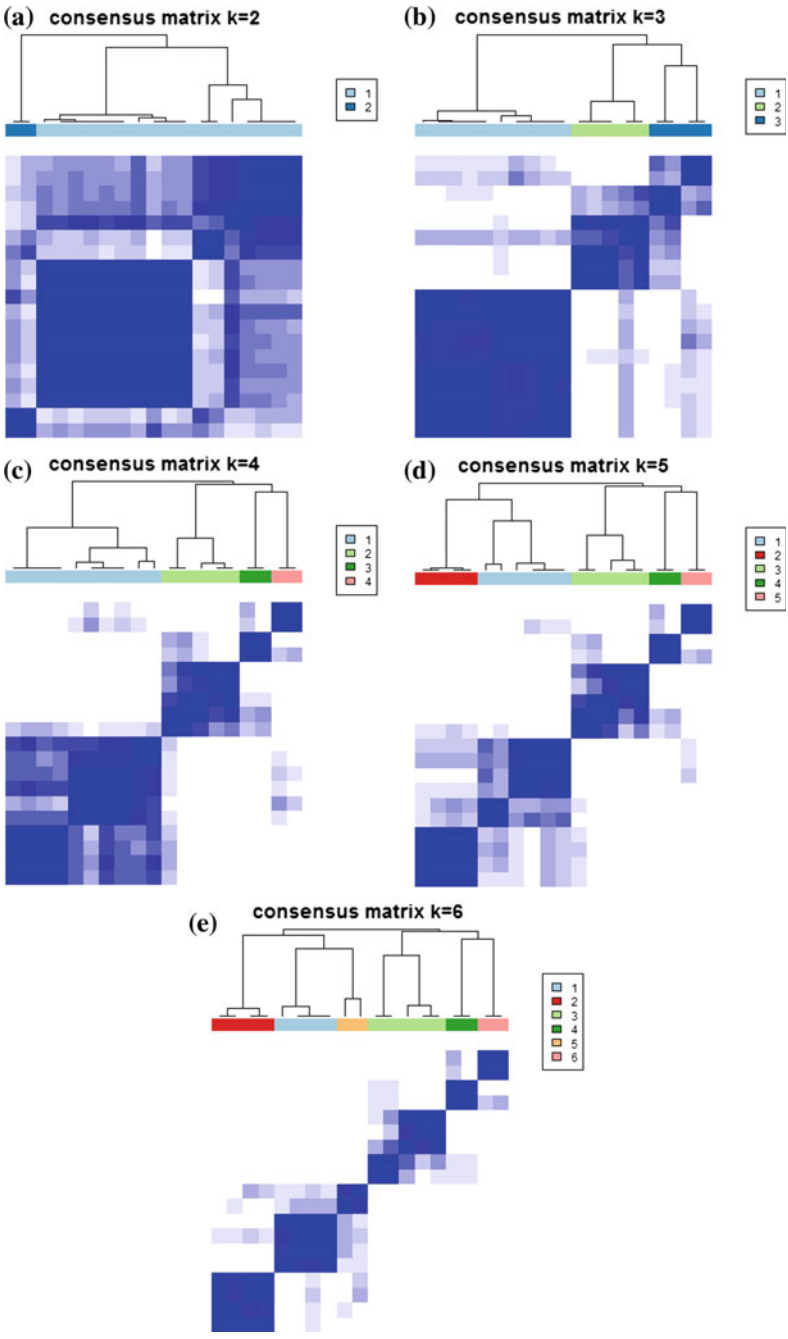

Fig. 6 Consensus matrix, **a** $k = 2$, **b** $k = 3$, **c** $k = 4$, **d** $k = 5$, and **e** $k = 6$

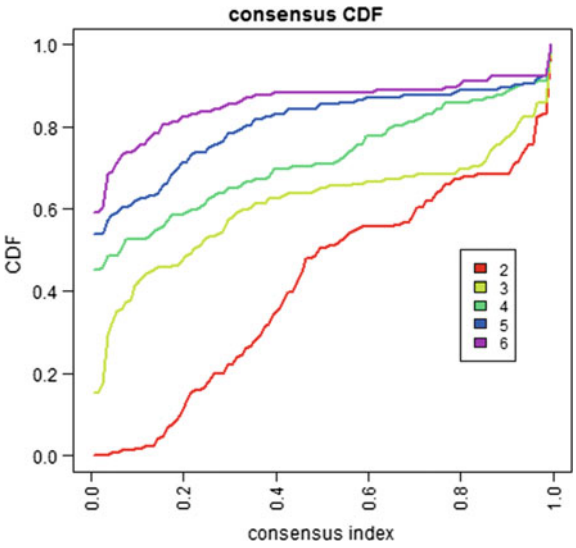

Fig. 7 CDF of consensus matrixes

Fig. 8 Cluster consensus

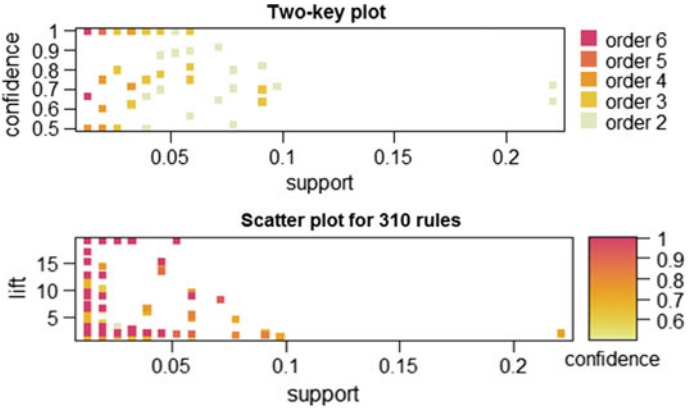

Fig. 9 Scatter plots of confidence and lift w.r.t support

5 Conclusion

This study has shown that the twitter data analysis with the use of different clustering techniques is beneficial. The same techniques can also be used in companies' stock market prediction and analysis and wherever Big Data analysis is required. In the analysis of social media datasets, we have concluded that TF-IDF finds its necessity in counting the important terms of a document. This analysis on twitter dataset gives the most efficient algorithm among the different algorithms mentioned in this work. The results depict that the consistency and efficiency of consensus clustering were better. *K*-means and *k*-medoid (PAM) produced almost same results. Hierarchical clustering is helpful only for short data, and it fails for large datasets. Overall, consensus results are satisfying. Hence, consensus clustering technique is best suited for any large dataset. The plot of consensus CDF graph, the probability of clustering of continuous data with respect to different clusters, esteems the accuracy in the clusters formed.

References

1. Bing, L.I., C.C.K. Chan, and, O.U. Carol. 2014. Public Sentiment Analysis in Twitter Data for Prediction of a Company's Stock Price Movements. In *2014 IEEE 11th International Conference on e-Business Engineering.*
2. Danyllo, W.A., V.B. Alisson, N.D. Alexandre, L.M.J. Moacir, B.P. Jansepetrus, and Roberto Felício Oliveira. 2013. Identifying Relevant Users and Groups in the Context of Credit Analysis Based on Data from Twitter. In *2013 IEEE Third International Conference on Cloud and Green Computing.*
3. Bahrainian, Seyed-Ali, and Andreas Dengel. 2013. Sentiment Analysis and Summarization of Twitter Data. In *2013 IEEE 16th International Conference on Computational Science and Engineering*.
4. Gokulakrishnan, Balakrishnan, Pavalanathan Priyanthan, Thiruchittampalam Ragavan, Nadarajah Prasath, and AShehan Perera. 2012. Opinion Mining and Sentiment Analysis on a Twitter Data Stream. In *The International Conference on Advances in ICT for Emerging Regions—ICTer 2012*, 182–188.
5. Bhuta, Sagar, Avit Doshi, Uehit Doshi, and Meera Narvekar. 2014. A Review of Techniques for Sentiment Analysis of Twitter Data. In *2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT).*
6. Kanakaraj, Monisha, and Ram Mohana Reddy Guddeti. 2015. NLP Based Sentiment Analysis on Twitter Data Using Ensemble Classifiers. In *2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN).*
7. Gautam, Geetika, and Divakar Yadav. 2015. Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis. In *2015 IEEE International Conference*.
8. Venugopalan, Manju, and Deepa Gupta. 2015. Exploring Sentiment Analysis on Twitter Data. In *2015 IEEE International Conference*.
9. Liu, C.-L., T.-H. Chang, and H.-H. Li. 2013. Clustering Documents with Labeled and Unlabeled Documents Using Fuzzy Semi-K-Means. *Fuzzy Sets and Systems* 221: 48–64.