Twitter sentiment analysis on demonetization using possibilistic fuzzy c-means approach

[1]Ms. Varsha S. Nagpurkar , [2]Ms. Jayashri Mittal
[1]Assistant Professor, [2]Assistant Professor
1Department of Computer Engineering,
[1]St. Francis Institute of Technology, Mumbai, India

Abstract : Social medias such as face-book, Google +, Instagram, Myspace, and Twitter are the collective of online communications channels dedicated to community-based input, interaction, content-sharing and collaboration, which helps people to share the public opinions about the products, political analysis and movie reviews, etc. In this paper, public opinion is created to make an attempt on DeMonetization (DeMon) policy in India. Initially, the DeMon twitter data were collected from November 9 to December 3, which was considered for experimental analysis. Usually, the noises such as URLs, positive and negative emoji, stop words were reduced from raw tweets in the pre-processing stage. The twitter sentiment analysis (TSA) was done by Possibilistic fuzzy c-means (PFCM) approach to pre-processed twitter data. The optimal clustering heads from the sentimental contents were found out by the proposed clustering method for acquired twitter data. The PFCM obtained the three types of results such as positive, negative and neutral to validate the TSA on DeMon. The experimental results showed that PFCM approach improved the accuracy in TSA upto 3-20% when compared with existing methods namely Bernoulli naive Bayes, multinomial naive Bayes, Support Vector Machine (SVM), decision tree and Linear Discriminant Analysis (LDA).

IndexTerms **- Demonetization, machine learning, possibilistic fuzzy c-means, twitter sentiment analysis.**

I. INTRODUCTION

Twitter is the vital platform used for communication and sharing information with friends [1]. Twitter makes the information easy to spread and read, because it allows the user to publish only 140 characters in a single tweet [2]. Several applications such as elections, reviews, sentiment analysis (SA) and marketing used Twitter as a blogging platform, because it provides a vast amount of data [3-4]. Generally, people tweets about various topics such as reviews on movies, products, brands, politics, etc on social media to share their opinions. [5]. This research work is mainly used to identify the public opinion on DeMon policy in India. The DeMon is one of the biggest political decision in India, which affected each and every citizen in India [6]. The DeMon is a step towards the dream of digital India, to tackle the concern of black money and to make India as a cashless economy [7]. Recent methodologies in TSA extracts the twitter text from online blogs, for classifying the text as positive or negative [8-9]. Challenges faced by the researchers in TSA are: neutral tweets are more common than positive and negative ones, which is very difficult to classify. Tweets are very short and often show limited sentiment cues [10].

Many researchers focused on the use of traditional classifiers, like naive Bayes, maximum entropy, and SVM to solve these problems [11-12]. In order to improve the classification accuracy, a new clustering classification methodology was implemented. In this experimental research, TSA was performed on the simulated DeMon twitter data. The proposed methodology consists of two phases such as, pre-processing and classification. The unwanted noise such as URLs, positive and negative emoji, stop-words was reduced by processing the twitter data which was considered as a first phase. The PFCM clustering approach was used for TSA on pre-processed twitter data. The advantages of both fuzzy c means and possibilistic c-means methods were combined for forming the PFCM. The sentimental contents of Twitter data were used for finding the optimal clustering heads by using the proposed clustering methodology. The PFCM obtained the results in three forms such as positive, negative and neutral. These twitter words were stored in the dictionary with an individual weight value, then the testing data was matched with the dictionary in order to evaluate the performance of the proposed approach.

This paper is organized as follows. Review of various recent papers on TSA is described in Section II. The proposed clustering methodology (PFCM) is presented for twitter classification in section III. Section IV shows the comparative experimental result for proposed and existing twitter sentiment strategies. The conclusion is made with future works in Section V.

II. LITERATURE REVIEW

Researchers have suggested several techniques for the TSA based on DeMon. In this scenario, a brief evaluation of some important contributions to the existing literatures are presented.

E. Kušen, and M. Strembeck, [13] evaluated a sentimental analysis at 2016 Austrian presidential election. In this literature, 343645 messages were extracted and analysed a data-set related to 2016 Austrian presidential elections. The developed methodology combined the methods like network science and SA and the clear polarization was found in terms of sentiments spread by twitter followers about the two presidential candidates. The approach was not able to recover the seven days old tweet because of application programming interface restriction. Hence, the approach lost the data permanently.

Jianqiang, Zhao, Gui Xiaolin, and Zhang Xuejun, [14] developed a word embeddings method based on large twitter data with the help of unsupervised learning by combining co-occurrence statistical characteristic and latent contextual semantic relationships between words in tweets. The sentiment features of tweets were formed by combining the word sentiment polarity score and n-gram features in word embeddings. The sentiment classification labels were predicted by feature set which was integrated into deep

convolution neural network (DCNN). The efficiency of word embedding method was validated by conducting experiments on five datasets when compared with existing techniques. The pre-trained word vectors used in DCNN had good performance in the task of TSA. While clustering the sentimental contents in large dataset, the computational time becomes a bit high.

M.Z. Asghar, F. M. Kundi, S. Ahmad, A. Khan, and F. Khan, [15] proposed a hybrid classification framework to overcome the issues of incorrect classification. The performance of twitter-based SA systems were improved by using four classifiers such as a slang classifier, a moticon classifier, the SentiWordNet (SWN) classifier, and an improved domain-specific classifier. The input text was passed through the first two classifiers such as emoticon and slag, after applying the preprocessing stage. In the final stage, SWN-based and domain-specific classifiers were applied to classify the text accurately. A limitation of the approach was the lack of automatic scoring of domain-specific words without performing a lookup operation in SWN, which may increase the classification accuracy.

N. Kannan, S. Sivasubramanian, M. Kaliappan, S. Vimal, and A. Suresh, [16] implemented a SA approach which analysed social media posts and extracted user's opinion in real-time. According to a selected set of hashtags related to a given topic, a polarity of a dynamic dictionary (DD) of words was constructed. The tweets were classified under several classes by new features which strongly fine-tune the polarity degree of a post that was introduced by the DD method. The approaches were validated by classifying the tweets which were related to US2016 Election. The experimental results showed that the detection of positive and negative classes with their sub-classes provided a good accuracy of the DD method. Still, the developed approach had a limitation that the automatic construction of DD using small samples.

I. El Alaoui, Y. Gahi, R. Messoussi, Y. Chaabi, A. Todoskoff, and A. Kobi, [17] proposed a Predictive Analysis on DeMon data using SVM (PAD-SVM). The system PAD-SVM involved three stages pre-processing, descriptive analysis and prescriptive analysis. The pre-processing stage includes cleaning the obtained data, performing and splitting the necessary data from the tweets. Whereas, the descriptive stage performed to find the most influential people and determining analytical functionalities. Moreover, the prescriptive analysis was performed to view the current mindset of the people and the society reacted to an issue in the current time. The model achieved more accuracy, which was used to override the existing data to find the predicted data. The method was tedious to proceed, because the developed method requires continuous prediction for calculating the polarity scores.

N.M. Dhanya, and U.C. Harish, [18] implemented various Machine Learning Algorithms (MLA) for classifying the Twitter data which gave an opinion about DeMon policy in India. The twitter data were collected from November 9th to December 3rd used for analysis. These collected data were pre-processed to remove the unwanted noises and missing values and gave input to MLA includes Navie Bayes, SVM, decision trees. The experimental results were carried out for analysing the performance of these approches and the obtained results stated that SVM showed better performance than other classifiers. The limitation of the MLA approach was reduction of accuracy rate, because the tweets in Hindi and Tamil language were not considered for classification (i.e. considered as neutral).

To overcome the above-mentioned issues and enhance the performance of TSA based on DeMon, a proposed clustering (PFCM) methodology is implemented in Section 3.

III. PROPOSED METHODOLOGY

TSA is the procedure of extracting the information from the huge amount of tweet data and then classifies the data into dissimilar classes named as sentiments. The important features from people opinions are mined by opinion mining known as SA. Previously, several MLA and statistical approaches are employed for TSA to extract the twitter data features. In this experimental analysis, a clustering based classification approach: PFCM is employed for analysing the people sentiments against DeMon. The proposed approach has two phases: training and testing phase, which is detailed below.

3.1 Training phase

In proposed methodology, the training phase contains four steps: data acquisition, pre-processing, classification and constructing dictionary. The working procedure of proposed methodology in training phase is given in figure 1.

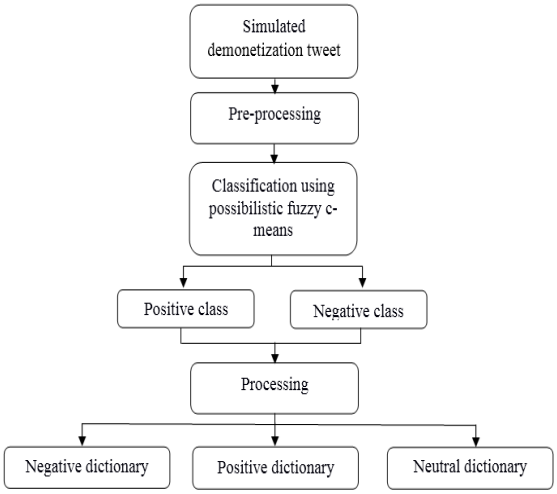

Fig.1 Working procedure of proposed methodology in training phase

3.1.1 Data collection and pre-processing

In the initial stage of TSA, the Twitter data were collected based on the specific hashtag (#DeMon). The stimulated database consists of raw tweets, dates, tweet Ids, senders and their locations. A total of 8,900 tweets were collected in the first phase, from November 9 to 16, 2016. Additionally, 3600 tweets were collected in the second phase, from November 17 to December 3, 2016. The sample tweet collection is given in figure 2.

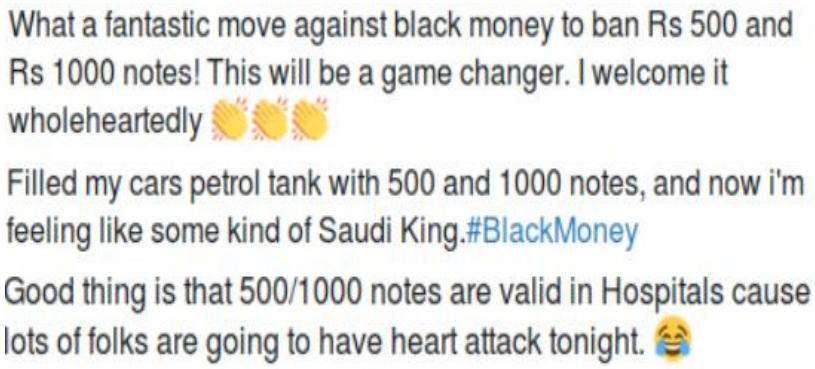

Fig.2 Sample tweets regarding DeMon

After the acquisition of DeMon twitter data, an important step in the TSA is pre-processing of acquired data. Before classifying the tweets as different classes such as positive, negative and neutral, the unwanted noises includes URLs, different emoji and stop words are removed from raw tweets. Then, an effective clustering based classification approach was employed for classifying the twitter classes. The brief description of clustering based classification technique is given below.

3.1.2 Classification using possibilistic fuzzy c-means

Each data point in the cluster is indicated by the membership grade in PFCM data clustering technique. The main aim of PFCM in clustering module is to reduce the Objective Function (OF), which is explained in the following equation (1).

(1)

With the following constraints as shown in equation (2) & (3):

(2)

(3)

Where, J_{PFCM} is the OF, U represented the partition matrix, T represents as the typicality matrix, V denotes the vector of cluster centres. The cluster centers and degree of membership are mathematically described in Equation. (4), which is the outcome of OF, achieved by using an iterative approach.

(4)

Where, data points are represented by n, number of cluster centers are described as c, which are described by the co-ordinates(x_j, v_i) and distance between data sets and cluster centers are calculated by using these co-ordinates.

For every cluster, the possibilities and memberships are constructed with cluster centers and normal prototypes in PFCM. Choosing the OF is an important aspect to acquire better clustering performance in cluster methodology. Whereas, the clustering performance depends on the OF, which is utilized for clustering. For developing an effective OF, the following set of requirements are considered.

- Distance between the clusters should be reduced.
- The data point distance should be reduced, which is allocated in the clusters.

The OF modeled the desirability between the data and clusters, which is further improved with driven prototype learning parameterα. The exponential separation strength between clusters highly depends on the learning procedure which is updated at every iteration. The parameter α is represented in equation (5).

(5)

Where, β is represented as sample variance, mathematically, it is represented in the equation (6).

(6)

Where,

Then, a weight parameter is introduced to calculate the common value ofα. Each point of the database consists of a weight in relationship with each cluster. So, the usage of weight function delivers a better classification outcome, especially in the case of noise data. The general equation of weight function is determined in the equation (7)

(7)

Where, w_{ji} is denoted as the weight function of the point j with the class i. Figure 3 represents the procedure of PFCM.

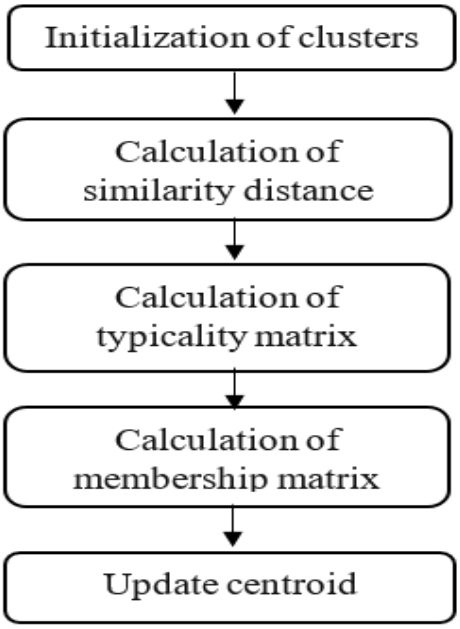

Fig.3 Process of PFCM

- **Initialization:** In every segment, the user-furnished the number of clusters, which are identical.
- **Calculation of similarity distance:** The distance between data points and centroids are evaluated when the number of clusters are identified for each segment.
- **Calculation of typicality matrix:** The evaluation of typicality matrix is obtained from the PFCM, after finding the distance matrix.
- **Calculation of membership matrix:** The evaluation of the membership matrix M_{ik} is performed by means of assessing the membership value of data point, which is obtained from the PFCM.
- **Update centroid:** The modernization of the centroids is updated, after generating the clusters.

When the modernized centres for every cluster becomes identical in successive iterations, then the above procedure reaches its final stage. The results obtained from the procedure can be identified as three forms such as positive, negative and neutral classes. Finally, assign a score to the sentiment words: 1 for positive words, 0 for neutral words and -1 for negative words. These twitter words are stored in the dictionary with an individual score value, then the testing twitter data is matched with the dictionary.

3.2 Testing phase

The testing phase of proposed TSA consists of four steps data acquisition, pre-processing, scoring and balancing and classification. The working procedure of proposed methodology in testing phase is represented in the figure 4.

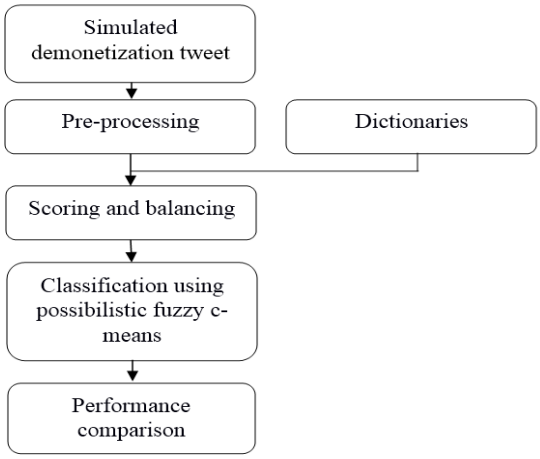

Fig.4 Working procedure of proposed methodology in testing phase

3.2.1 Scoring and balancing

After the acquisition of DeMon tweets, pre-processing is carried out by cleaning the obtained twitter data, execution of missing value treatment and eliminating the unnecessary emoji from the tweets. Normally, the language used in the social media posts is not conventional, which contains some special words such as words written in upper case and repetition of the words more than two consecutive letters. In most of SA approaches, these kinds of words are not fully exploited, whereas the polarity degree of the post could be decreased or intensify by extended words. While relying on these metrics, the PFCM balance the word score by concentrating more on these kinds of words. To highlight the emphasize sentiment, the PFCM employed the upper and extended case words. Here, a weight value "+1" is added to the positive words, "-1" is added to the negative words and "0" is added to the neutral value. For instance, if score (hate) =1, then score (haaate) =score (HATE) =score (HAAATE) =2.

3.2.2. Determine the polarity degree

The polarity of a DeMon tweet is determined by adding the independent score values of words t, and m represented the length of t. The general formula to calculate the polarity degree is determined in the equation (8).

(8)

Based on the score value, the DeMon tweets are classified into three classes (negative, positive and neutral class) using PFCM technique. In order to evaluate the proposed methodology performance, the testing twitter data are coordinated with the dictionary (contains trained twitter data).

IV. EXPERIMENTAL RESULT AND DISCUSSION

For experimental simulation, Hadoop software was employed on PC with 3.2 GHz with i5 processor. In order to estimate the efficiency of PFCM algorithm, the performance of the proposed PFCM method was compared with Bernoulli naive Bayes, multinomial naive Bayes, SVM, decision tree and LDA classification methods [18] on the simulated DeMon twitter data. The performance of the proposed methodology was compared by means of accuracy, precision, recall, and f-measure.

4.1 Performance measure

The performance measure is defined as the relationship between the input and output variables of a system is understood by employing the suitable performance metrics like precision and recall. The general formula for calculating the precision and recall of the TSA is given in the equations (9) and (10).

(9)

(10)

Accuracy is the measure of statistical variability and a description of random errors. The general formula of accuracy for determining the TSA is given in the equation (11).

(11)

Where, TP represents as true positive, FP denotes false negative, TN represents true negative and FN is false negative. F-measure is the measure of accuracy test and it considers the both precision P and recall R of the test in order to calculate the score. The general formula for F-measure is given in the Equation (12).

(12)

4.2 Experimental analysis on acquired DeMon twitter data

In this experimental research, simulated DeMon twitter data is used for comparing the performance evaluation of existing methodologies and the proposed approach. In table 1, the precision and recall value of proposed and existing methodologies are compared for three classes: positive, neutral and negative. The average precision value of the proposed technique: PFCM delivered 0.92 and the existing methodologies: Bernoulli naive Bayes, multinomial naive Bayes, SVM, LDA and decision tree delivered 0.74, 0.78, 0.87, 0.74 and 0.78 of precision. Similarly, the average recall value of the proposed technique delivered 0.88 and the existing methodologies delivered 0.67, 0.63, 0.84, 0.62, and 0.63 of recall. The graphical representation of an average precision and recall is denoted in the figures 5 and 6.

Table 1 Proposed and existing methodologies comparison by means of precision and recall

Methodologies	**Precision**				**Recall**			
	Positive	**Negative**	**Neutral**	**Average**	**Positive**	**Negative**	**Neutral**	**Average**
Bernoulli naive Bayes [18]	0.83	0.90	0.64	0.74	0.26	0.24	0.99	0.67
Multinomial naive Bayes [18]	1	1	0.61	0.78	0.12	0.19	1	0.63
SVM [18]	0.95	1	0.79	0.87	0.84	0.59	1	0.84
LDA [18]	0.89	0.32	0.81	0.74	0.41	0.68	0.70	0.62
Decision tree [18]	1	1	0.61	0.78	0.12	0.19	1	0.63
Proposed (PFCM)	**0.98**	**1**	**0.86**	**0.92**	**0.89**	**0.76**	**1**	**0.88**

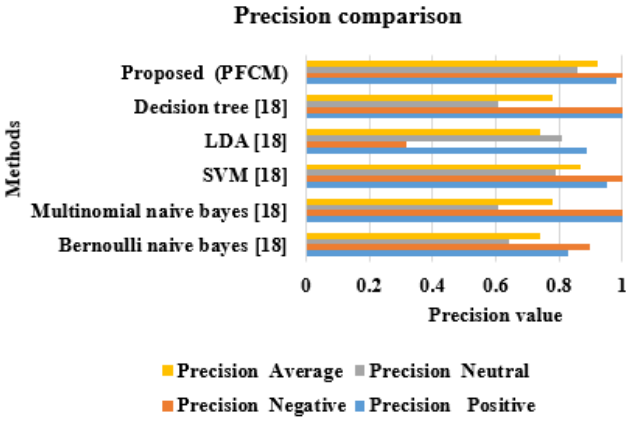

Fig.5 Graphical representation of precision comparison

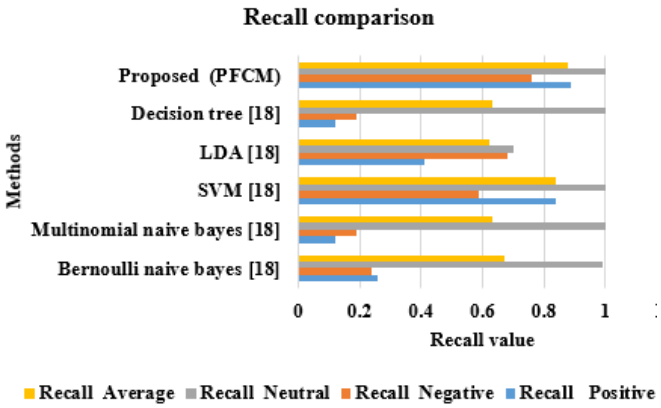

Fig.6 Graphical representation of recall comparison

In table 2, the validation result shows that the proposed methodology outperformed the existing methodologies in terms of f-measure and accuracy. The average f-measure value of the proposed technique: PFCM delivered 0.86 and the existing methodologies: Bernoulli naive Bayes, multinomial naive Bayes, SVM, LDA and decision tree delivered 0.61, 0.54, 0.83, 0.64 and 0.54 of f-measure value. Similarly, the accuracy of the proposed methodology delivered 87.76% and the existing methodologies delivered 66.75%, 63.25%, 84%, 62.25% and 63.25% of accuracy. The graphical representation of average f-measure and accuracy is represented in figures 7 and 8.

Table 2 Proposed and existing methodologies comparison by means of f-measure and accuracy

Methodologies	**F-measure**				
	Positive	**Negative**	**Neutral**	**Average**	**Accuracy (%)**
Bernoulli naive Bayes [18]	0.39	0.38	0.78	0.61	66.75%
Multinomial naive Bayes [18]	0.22	0.31	0.76	0.54	63.25%
SVM [18]	0.83	0.74	0.88	0.83	84%
LDA [18]	0.56	0.44	0.75	0.64	62.25%
Decision tree [18]	0.22	0.31	0.76	0.54	63.25%
Proposed (PFCM)	**0.85**	**0.78**	**0.92**	**0.86**	**87.76%**

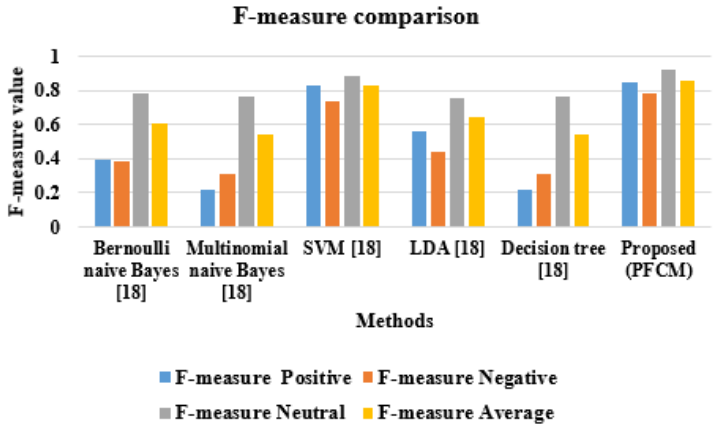

Fig.7 Graphical representation of f-measure comparison

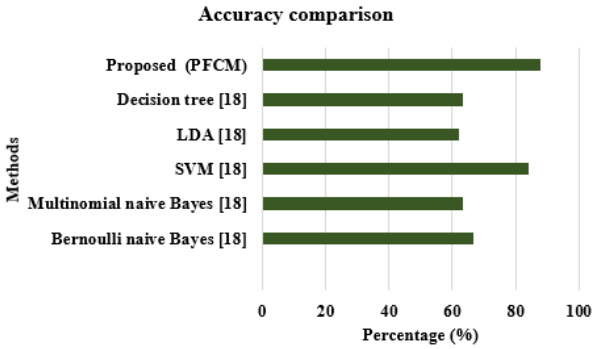

Fig.8 Graphical representation of accuracy comparison

Table 1 and 2 clearly shows that the proposed PFCM approach improved the classification accuracy in TSA up to 19% compared to the existing methods in the simulated DeMon twitter data. In this scenario, clustering based classification process is performed to predict the best results. The clustering based model for twitter sentiment helped to improve the sentiment classification accuracy. Compared to the existing schemes, the proposed method works effectively in terms of precision, recall, accuracy and f-measure.

V. CONCLUSION

TSA is one of the emerging research fields for analysing and identifying the sentiments and viewpoints of users. This research work analysed the effects of DeMon policy implemented by the Indian government using sentimental analysis concept. In this experimental research, the DeMon twitter data were collected from the time period of November 9 to December 3, 2016. The acquired twitter data was pre-processed by eliminating the unnecessary emoji from the tweets, and the execution of missing value treatment. The pre-processed data was utilized for TSA using PFCM clustering methodology. This clustering methodology identified the optimal cluster heads from the sentiment contents. The experimental investigation of PFCM was verified on simulated DeMon twitter data, which showed the superiority of the proposed approach. The classification rate of DeMon twitter data is better in the proposed methodology than the previous methodologies. Compared to other existing approaches in TSA, the proposed scheme delivered an effective performance by means of accuracy. The developed approach improved the classification accuracy of around 3-19% compared to the previous methods. In future work, to improve the classification rate, a new multi-objective classification approach will be developed.

REFERENCES

[1] Luis, T. and Mancera, J. 2019. Dynamic profiles using sentiment analysis and twitter data for voting advice applications. Government Information Quarterly.

[2] Shirdastian, H., Laroche, M. and Richard, M.O. Using big data analytics to study brand authenticity sentiments: The case of Starbucks on Twitter. International Journal of Information Management, 2017.

[3] Himja, K. and Sahu, S.K. 2018. Bat inspired sentiment analysis of Twitter data. Progress in Advanced Computing and Intelligent Engineering. Springer, 639-650.

[4] Daniel, M., Neves, R.F. and Horta, N. 2017. Company event popularity for financial markets using Twitter and sentiment analysis. Expert Systems with Applications, 71: 111-124.

[5] Ruan, Y., Durresi, A. and Alfantoukh, L. 2018. Using Twitter trust network for stock market analysis. Knowledge-Based Systems, 145: 207-218.

[6] Arun, K., Srinagesh, A. and Ramesh, M. 2017. Twitter Sentiment Analysis on Demonetization tweets in India Using R language. International Journal of Computer Engineering in Research Trends, 4(6): 252-258.

[7] Singh, P., Sawhney, R.S. and Kahlon, K.S. 2017.Sentiment analysis of demonetization of 500 & 1000 rupee banknotes by Indian government. ICT Express.

[8] LaBrie, R.C., Steinke, G.H., Li, X. and Cazier, J.A. 2017. Big data analytics sentiment: US-China reaction to data collection by business and government. Technological Forecasting and Social Change.

[9] Komorowski, M., Do Huu, T. and Deligiannis, N. 2018. Twitter data analysis for studying communities of practice in the media industry. Telematics and Informatics, 35(1): 95-212.

[10] Vyas, V. and Uma, V. 2018. An Extensive study of Sentiment Analysis tools and Binary Classification of tweets using Rapid Miner. Procedia Computer Science, 125: 329-335.

[11] Singh, T. and Kumari, M. 2016. Role of text pre-processing in twitter sentiment analysis. Procedia Computer Science, 89: 549-554.

[12] Birjali, M., Beni-Hssane, A. and Erritali, M. 2017. Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks. Procedia Computer Science, 113: 65-72.

[13] Kušen, E. and Strembeck, M. 2018. Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections. Online Social Networks and Media, 5: 37-50.

[14] Jianqiang, Z., Xiaolin, G. and Xuejun, Z. 2018. Deep convolution neural networks for Twitter sentiment analysis. IEEE Access, 6: 23253-23260.

[15] Asghar, M.Z., Kundi, F.M., Ahmad, S., Khan, A. and Khan, F. 2018. T-SAF: Twitter sentiment analysis framework using a hybrid classification scheme. Expert Systems, 35(1).

[16] Kannan, N., Sivasubramanian, S., Kaliappan, M., Vimal, S. and Suresh, A. 2018. Predictive big data analytic on demonetization data using support vector machine. Cluster Computing, 1-12.

[17] El Alaoui, I., Gahi, Y., Messoussi, R., Chaabi, Y., Todoskoff, A. and Kobi, A. 2018. A novel adaptable approach for sentiment analysis on big social data. Journal of Big Data, 5(1): 12.

[18] Dhanya, N.M. and Harish, U.C. 2018. Sentiment analysis of twitter data on demonetization using machine learning techniques. In Computational Vision and Bio Inspired Computing Springer, Cham, 227-237.