Journal of Engineering Science and Technology
Vol. 14, No. 2 (2019) 922 - 934

ADDRESSING SOCIAL POPULARITY IN TWITTER DATA USING DRIFT DETECTION TECHNIQUE

LEENA A. DESHPANDE*, M. R. NARASINGARAO

KLEF Deemed to be University, Green Fields, Guntur District,
Vaddeswaram, Andhra Pradesh 522502, India
*Corresponding Author: deshpande.leena27@gmail.com

Abstract

In the social networking site, the use of Twitter data is popularly used by web users. People are using twitter sites to express their opinion. Due to this wide usage, data mining techniques are used for further prediction. A traditional classification model is used to classify the data based on past-labelled data. However, in textual data application, data arises with fluctuating patterns and size, which leads to new features and drift, which reduces the accuracy in the prediction. Such a pattern usually found in the user's social media post like twitter. This inspires us to develop and analyse a method, which identifies a deviation in the data and retrain the model according to drift identified. The weight-based features and *n*-Gram technique are used to identify the unknown labels in the text. It further detects a drift in the data. Further, a cause of the drift is analysed by finding new features evolved in the concept. Our experiment shows that that pre-processing and drift detection techniques significantly improve the classification accuracy. Proposed work may help the government to analyses the awareness of people living in India for the welfare of society.

Keywords: Classification, Concept drift, Data pre-processing, Drift, Term frequency.

1. Introduction

To understand the sentiments of people living in India, sentiment analysis, opinion mining is primarily used to determine whether the policies are really reaching to people and how effectively it has been accepted from user's perspective. Social media network plays a vital role of spreading the awareness among the citizen of the country, it is very important to analyse whether the schemes or policies launched by the government of India are reaching out to the people or not. In India, e.g., a policy like Swachh Bharat Mission, also known as "Clean India" (2014), was successful all over the region, while another policy like Bachat Lamp Yojana (2009), was not popular like the one stated earlier [1, 2]. Likewise, various other policies need to be analysed for predicting its impact on the people. By analysing various pattern about user's interest and response, such policies can be improvised using people's feedback. Social Media platforms have become a popular medium for users to share information. This inspires us to develop a utility to watch on the social postings of their schemes.

However, the textual contents in social media are highly unstructured, informal, and contains a different writing style. In twitter text, a number of words, messages may be written for a very short time period. Users opinion may get changed after a certain amount of time, or there might be a sudden change in the topic and become viral over the social media site. Due to such a varied pattern, there is a change of context over a period of time, which is called drift. For an example, on Twitter, people were talking of Swatch Bharat (Clean India) for some time and then suddenly started talking about demonetization as it suddenly picked up the user's interest due to a sudden step of demonetizing 500 and 1000 rupee note to cut down the illegal activity and terrorism. The sudden announcement of demonetization become a sharp criticism among the people, over the social media network, TV news, etc. This causes a drift in data where the user's regular usage pattern of social posts deviated. Such type of data requires retraining of a model for accurate prediction. A classification model is proposed to detect a drift in the data distribution and my various patterns out of this varied data

In text data mining, extracting complex and relevant features from the text, and further selecting a classification algorithm is an important and fundamental task. Our approach is to overcome the problem of misclassification by identifying correct labels and drift in the data.

In this paper, the authors try to investigate the following questions.

- Does the textual preprocessing identify new class labels for accurate classification?
- Does a change in the popularity of data causes due to change in the concepts arrived in the data?
- Can a drift be identified and how an adaptive technique improves the overall accuracy of a classifier?

The contribution of this paper is a new approach for identifying unclassified class labels by retraining the class models by finding new class labels and concept to improve the overall classification accuracy. With our approach, the polarity of sentiments by applying weighted labels for extracting new features. With our approach, semantically most useful labels are given higher weightage and classifier is retrained for drift detection.

The goal of our proposed system is to,

- Design and develop a classification-based model for detecting drifts in data streams
- Automatically label the data from a set of labelled documents
- Detect and analyse the drift in the data

The next section of this paper explores the existing methodologies for handling concept drift. Our proposed approach in the next section experiments the dataset taken from twitters containing the posts given on government schemes, which are implemented by the Indian government for the people. Our evaluation methods and results are discussed in this section.

2. Problem Formulation

To improve the performance of the classifier, researchers have proposed a solution in various broad areas like pre-processing and feature extraction, drift detection techniques, novelty detection and concept drift techniques concept drift has introduced new challenges in the classification of the textual data stream. Several classification algorithms have been evolved to deal with concept drift in data streams.

However, they mainly focused on one of the concepts drifts: sudden, gradual or recurring [3, 4]. The gradual type of drift is associated with a slower rate of changes in data streams. Typically, the data instances from different source distributions start mixing, where new source distribution increases and that of old distribution decreases over time. When a huge chunk of data is generated over the web memory requirement and storage space are limited. For handling these challenges, the most popular sliding window processing model [5, 6] is used. For the Twitter data stream, weighted parameters are extensively used for labelling and classifying the data.

For pre-processing the text data, Stanford Parser Tools and lemmatization NLTK toolkit was used for POS tagging, parsing and removing stop words. Masud et al. [7] proposed that a concept drift evolves as a conditional change due to a change in decision boundaries. However, their work is restricted to the standard UCI repository of the medical dataset. Rather than relying on single classifier researchers also applied an Ensemble technique for detecting drift. Accuracy Updated Ensemble (AUE2) algorithm works well on different types of drifts [5]. The system maintains a weighted group of base classifiers and predicts the class of incoming data instance by combining predictions of base classifiers by weighted voting rule.

Researchers also proposed clustering techniques, which proved better in classifier accuracy. A decision tree for generating clusters of data streams is used and a drift is detected using deviation of data instance from given clusters. The ensemble of both classifiers and clusters are built using training data and clusters of testing data. A concept drift is detected using unlabelled instances to form a separate cluster. Zhao et al. [8] proposed a semantic network of user's tweet for detecting rise and fall of multiple dimensions over time. Ke et al. [9] explained that the accuracy of recommendation is improved by addressing the social influence of users by weighing the user interaction according to time distance. A matrix

factorization technique is used as a temporal dynamic for identifying social relationships. In novelty detection, use of old weights and new weights helps to maintain high generalization of old and new concepts and provides high accuracy, as information in old concept is used for learning new concept [6, 10].

The drift detector works for unbalanced, high dimensional data set with any proportion of block size. According to Escovedo et al. [11], a dynamic block size has a variable impact on classification accuracy. Traditional methods classify the instances, which are trained, however, due to change in data distribution, classifier may misclassify the data as it is not trained for the drift data, prediction accuracy may get a decrease in the presence of drift, supervised algorithm degrades its performance, unless a new class has been trained with labelled instances. Novelty detection is based on the type of distribution of data and features evolved. The feature-based techniques are used to weigh the important features and classify data. It includes basic lexicon based and linguistic-based methods.

Based on studies by Littlestone and Warmuth [12] and Amarnath and Balamurugan [13], a pool of algorithm is used and then weighted majority algorithm is applied to correct the mistake to improve the prediction. Weights are updated for wrong inputs and recursively applied to improve the accuracy. Best algorithm from the pool is chosen Drift or variation also lies in the type of data generated leads to complex hidden pattern Incremental or batch approach have been proposed by many researchers for classification problem A large scale streaming classification problem is solved using Ensemble classifier Incremental ensembles have been built [14]. Whereas, spatiotemporal data also holds a large amount of complex and hidden knowledge. Masud et al. [15] commented that the analysis of such data using tweets can be used for detection of spatial trajectories of moving objects.

Oza and Russel [16] and Parker et al. [17] stated that the other approach deals with the problem of concept evolution or novelty detection, in addition to concept drift. The identified trajectories help the analyst to find interesting patterns, as well as drift, occurred in patterns. For example, according to Senaratne et al. [18], the work is based on the famous singer, Lady Gaga's tweets. These tweets are used for detection of spatial trajectories of her world tour. Authors provided a framework that utilizes kernel density estimation for detecting hotpot clusters of twitter activities.

Concept evolution occurs when new classes are introduced in data streams. A k-NN based classifier is proposed by Masud et al. [19], which detect the concept evolution problem by proposing an adaptive threshold for detecting outliers. Singh and Kumari [20] explained that the analyses insignificance and significance of slang word used in 2014 election held in India, author detects a concept drift using sliding window protocol for finding trending topics to predict election results. Another twitter-based classification model is proposed by Lifna and Vijayalakshmi [21], which is again prediction model of 2014 election data.

An unbalanced data set we have proposed a machine learning and lexicon-based approach, which pre-process and relabel the tweets using weight-based measures. The trending topics in data streams are populated based on the frequency of the topic. A windowing technique is used on Indian election data set to identify drift. In our case, a novelty is detected after pre-processing and classifying tweets. Using machine-learning approach, we found that outlier and noise in the data are given to

the classifier after a certain interval of time. We cannot treat such data as outlier or noise and applied an n-gram technique on those to get the novelty class.

3. Problem Solution and Discussion

The proposed system for handling concept drifts in data streams uses a pre-processing technique, which categorizes labels. We have fixed the labels to three categories 1. *SwachhaBharat (SWC)* ("Clean India"), 2 *Make in India (MKI)* and 3. *Demonetization (DEM).* It builds the base classifiers and retains the system when drift occurs. In text data, people might change their writing style or might shift to new interest and started posting newly introduced subject, which might not be of their interest before.

Our work identifies such fluctuations in the data. To identify the drift, three basic parameters are used, the accuracy of classifier, the weight assigned to tweet for detecting the potential of the tweet and error rate of the classifier. As the original classifier gradually decreases its accuracy, a new classifier is rebuilt by changing the window size as soon as drift deviates. When the block of the fixed number of data instances is formed, the block-based classifier is developed. The classification of the incoming data instance is done using a weighted majority of base classifiers using weighting functions [9]. The proposed system is divided into 3 major steps:

Step 1

Weight-Based Classification: In this, each document is given weights as per the words in the documents these weights are assigned based on class labels and term frequency of the word in the dictionary. After particular class labels get a maximum weight then the document is classified to that class label.

$$W_i(t) = 1/(1 - A_i(t)) \tag{1}$$

where, $W_i(t)$ is the weight of base classifier C_i at time t and $A_i(t)$ is the accuracy of classifier C_i at time t.

Step 2

The accuracy and error rates are monitored for each type of classifier continuously over blocks of data instances using error rate function as:

$$Eij = \left(1 - f_{iy}(x)\right)^2 \tag{2}$$

where Eij is the error rate of classifier C_i on recent block B_j of data instances $f_{iy}(x)$ and is the probability given by the classifier C_i that x is an instance of class y.

Step 3

Drift detection is done when the error rate Eij of classifier Ci over block Bj exceeds user-defined threshold value. Drift is detected as the value of error rate monitoring crosses a certain threshold. These drifts are analysed and the ensemble is updated accordingly.

In the proposed system, the pre-processed data or bigrams are modified at each stage. Bigrams (and n-Gram) are used in training data. They are also used in feature extraction of the data. Each word is given a weight according to the word in the document. These weights are assigned based on term frequency occurrence in each

tweet. We have categorized three class labels; a tweet will be classified into a class label having maximum weight age. Removal of Unicode and Special character Conversion from *"json"* to *".csv*" is performed as part of cleaning data. Further, features selection is done and a dictionary is formed on the basis of unigram calculation of words in a document and a term-document matrix is formed. Following major parameters are derived for labelling the tweets. Date of creation, time, Tweet, number of Retweets, followers, follows, location, popularity index.

Dictionary creation: Using a lexicon approach, we have created our own dictionary with predefined words list. For example, politics dictionary will contain predefined standard words as Prime Minister, election, assembly, votes, country and others. Likewise, each word in the dictionary has assigned a weight based on how many times a word appears in the tweet text. Dictionary is updatable since we are training the data in each pass and we set the threshold value for the number of occurrences of the word. Words occurring with a high occurrence, medium occurrence, lower medium occurrence and low occurrence will gain weight as 4, 3, 2 and 1. It will be applied only after stopping word removal and stemming. Therefore, the value for each tweet will be based on the tweet value, followers, follows, retweet, and location.

$$\text{Tweet Weight} = \sum(\text{parameter} * \text{parameter\_val}) \quad (3)$$

$$\sum_1^n pw * Pv \quad (4)$$

where p_w and p_v are parameter weight and parameter values respectively.

If we have a tweet: "Government is launching a new Swachhbharat campaign." and we have a dictionary of terms along with weights (shown beside) as follows:

Swachhbharat→4
Clean→(1),
Campaign→(1),
Progress→(1)
So, the weight assigned to the tweet will be:

*$Total\ weigt$ = [1*4(Swachhbharat)]+ [0 *1(clean)] + [1* 1(campaign)] + [0 * (progress)] =5*

PI is calculated for the tweet, which a ratio of total weight calculated with the whole compared to each term in the class label dictionary. Therefore, if there are multiple matching dictionary terms of same label/category in the particular tweet then the weights are simply added to get the total weight of that tweet with respect to the dictionary.

3.1. Concept evolution

The concept of evolution is handled by creating a new class based on results. Out of around 40,000 tweets, it is found that over 3000 tweets were categorized as "other". Other category is the outlier class and not the noise. A list of potential keywords and their relations is found out from this category in order to detect a drift in the data stream. The key idea of relabelling the tweets is to identify labels, which are getting old and reassigning the weights to the keywords. Classifiers are weighted according to their accuracy after each incoming example. With each

misclassified instance (in our case the keywords) weight of classifier that misclassified the data instance is decremented by a certain predefined value.

At regular intervals, the accuracy of the whole ensemble is tracked to know if any classifier needs to be removed or added to the ensemble. In the proposed methodology, a concept evolution is based on various factors as, chunk size (windowing technique), random selection of sample instances, weight-based features in tweets, and appropriate classifiers.

3.2. Experimental setup

In this section, the experimental study proves that the objectives, which were mentioned in the introduction section met with the approach of concept drift detection. The product runs on Windows and Linux-based Operating systems. For data collection, Twitter API with the help of *R* tool is used. Tweets are extracted daily. More than 40,000 tweets are extracted from 12th October 2016 to 2nd November 2016. The sample of tweets is shown in Table 1.

Table 1. Sample tweets extracted.

Tweet (date/time)	Tweet text
10/13/2016 2:46 AM	What an incredible day! So fulfilling to watch young faces light up. Proud SANA could give them drinking water
10/13/2016 3:44 AM	An awareness campaign was organized by Bhilai Nagar Nigam today. (1) #MyCleanIndia #SwachhBharat https://t.co/NmSBvWdl8l
10/13/2016 8:46 AM	RT @SwachhBharatGov: An awareness campaign was organized by Bhilai Nagar Nigam today. (1) #MyCleanIndia #SwachhBharat https://t.co/NmSBvWdlâ€¦
10/13/2016 11:44 AM	@OfficeOfRG @divyaspandana @CatchNews #SwachhBharat kya hua gandee log gandee soch https://t.co/e9sVIx1s2h
10/13/2016 14:35 AM	Citizens were educated on the hazards of Open Defecation. (2) #MyCleanIndia #SwachhBharat https://t.co/vLpglZfMqD
10/13/2016 14:36 AM	The Corporation's #Swachhata mascot was used as a part of the rally to educate people. (3) #MyCleanIndiaâ€¦ https://t.co/PBLzEH2fRR
10/13/2016 14:39 AM	The rally was conducted as a part of the Nagar Nigam's #SwachhataPakhwada initiative. (4) #MyCleanIndiaâ€¦ https://t.co/2rYxjeoxnI

3.2.1. Pre-processing steps

For pre-processing of tweets Python as well as *R* script is written. Removal of Unicode, special characters, repetitive words and special numbers are removed using *R* script. Special and common Hindi language words are removed using a python script.

Words, which do not contribute in predictions are removed like ‘Kya’, “hum’, Tum’, ‘dekh’, etc. Tweets are collected with #tag of labels and also on popular words related to the label. Excel’s find and replace function is used to make some of the attribute unified, e.g., in the dataset, an attribute LOCATION has different spellings of the cities, like Kolkatta or Calcutta, Chennai or Madras and so on.

A unique name is given to a similar location in such cases. Table 2 summarizes total no of tweets actually extracted from twitter data.

Table 2. Number of tweets extracted in three different categories.

Tweet dates	Category (policy)	Number of tweets
12 /10/2016 to 2/11/2016	Swachha Bharat(SWC)	9609
	Make in India (MKI)	3451
	Demonetization (DEM)	4288
	Other	17631

3.2.2. Polarity score for popularity index

The popularity of tweet is calculated as low, medium and high based on the weighted matrix. After retraining the data, using N-gram techniques, the dictionary of "Swachh Bharat", "Demonetization" and "Make in India" category is refined with new keywords. The rest of other Tweets are considered as noise as they do not fall within the threshold value. The new N-gram Keywords, which gain higher weights among all three categories is labelled into the respective category. A popularity index is calculated, as shown in Table 3 is based on the age of the tweet, the weights assigned, the retweet count and the location. For example, if the location is from the popularly dense capital city like Delhi, Mumbai, etc., a higher weight will be assigned to the location. Age of the tweet gets reduced when tweet become old. After a block of 20 tweets, the weight assigned to that tweet is reduced. Based on these parameters a tweet score is calculated.

However, a new class altogether has arisen, which identifies a concept of evolution. Table 4 portrays about Prime Minister Modi (PM Modi) class, which was not labelled before in the three predefined categories. Rest all tweets are taken as noise so removed from the corpus. To assess the accuracy and weights applied, popular classification algorithm, Naïve Bayes is used in the proposed methodology. The existing data set is unbalanced data and hence, new class labels are discovered when accuracy degrades below 70%. A classifier is re-built using new keywords and output is tested. To compare the obtained result, the accuracy and error rates are monitored for each stream of the block (of size 50, 100, 150,) using the Error Rate function as shown in Eq. (2). A drift is detected based on the error rate value. Error rate value is continuously monitored and when it reaches, a certain threshold value drift is detected.

Table 3. Overall weight measure.

No.	Weights	Popularity Index
1	> 0 and < 0.4	Low
2	>0.4 and < 1.0	Medium
3	>1.0 and < 7.0	High

Table 4. Popularity of tweets (low, medium, high).

Category (policy)	Labelled tweets	Low	Med	High
Swachha Bharat (SWC)	10,004	2323	6636	5097
Make in India (MKI)	3478	630	403	554
Demonetization (DEM)	4312	552	88	640
PM Modi (New class)	11,189	-	-	-

In order to find out the evolution of the new concept, popularity measure is to be assessed like how people's interest in moving from one policy to another based on popularity value. However, to know this gradual change, popularity is further derived into low, medium and high

Figure 1 indicates that people have shown more interest in "Swachh Bharat "(Clean India) scheme than demonetization though demonetization is a sudden and major step taken by the government of India, it has received comparatively lower popularity.

To determine the change in the class label, change detection is carried out using the sliding window approach. In this approach, a time window is set, where date wise tweets are taken in a fixed size window. Change is identified as soon as it detects. A threshold is set to consider a change in the data stream. A drift is calculated with varying window size. (window size = 50, 100,200 and 500) as shown in Figs. 2 to 4.

It is clear from Fig. 2, that a stable drift is determined at window size 50. A drift is detected after 29th October and remains constant for a longer period. Thus, data is changed exponentially for a certain period of time, remain stable and again changes exponentially.

Choosing a correct window size proves a stable performance even in the presence of concept drift. Above graph shows the trending topic for "demonetization "with window size 50 and 500. A drift is detected for varying window size (starting from size 10), here the only graph of size 50,100 are shown.

It is clear that classifier performance deteriorates rapidly with a change in window size. Choosing a correct window size proves a stable performance even in the presence of concept drift. Thus, to label a varied pattern in the arriving data, a windowing technique is useful.

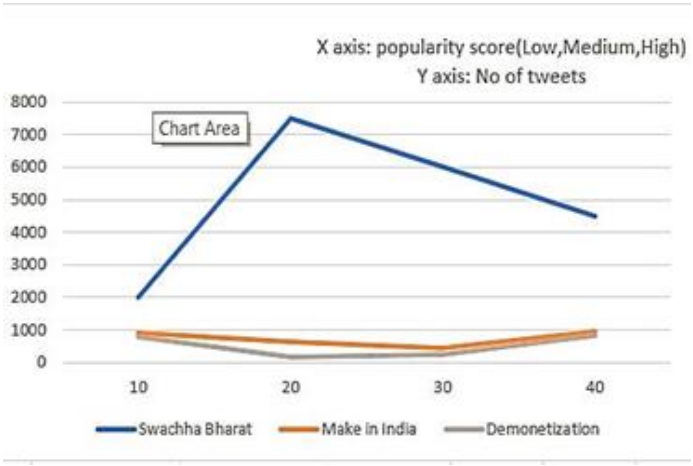

Fig. 1. Classification of tweets.

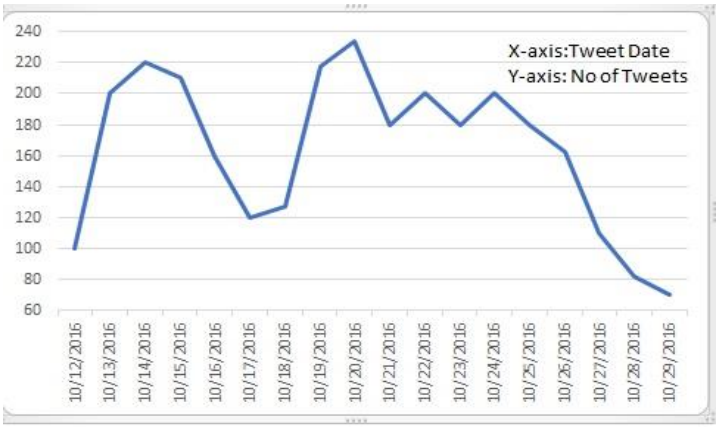

Fig. 2. Concept-drift in twitter trending topics "swachha" (window-size = 50).

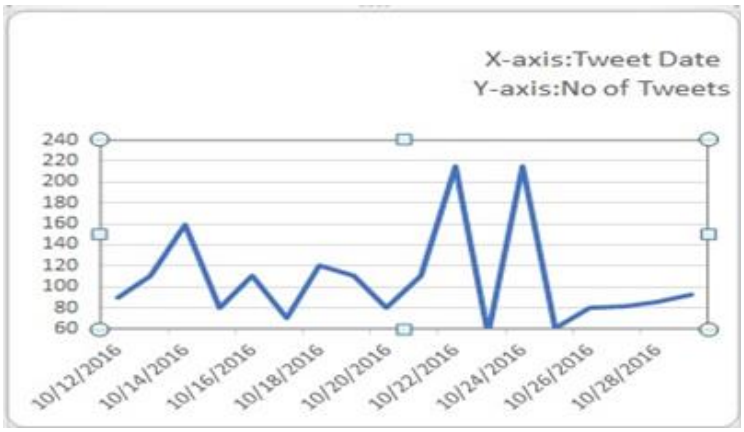

Fig. 3. Concept-drift trending (window-size = 50).

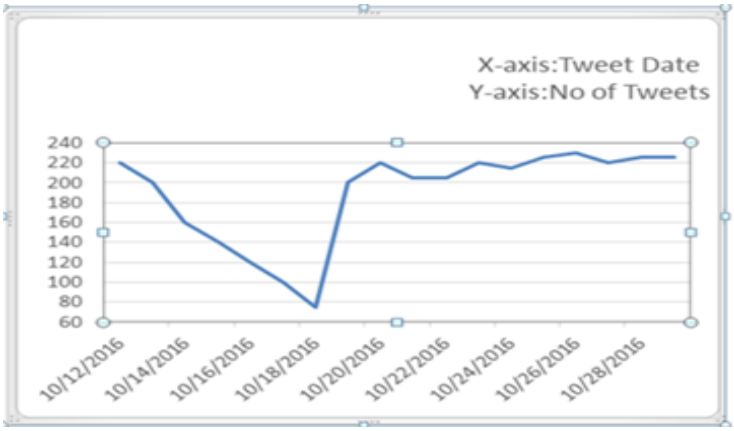

Fig. 4. Concept-drift trending (window size = 500).

4. Conclusions

Finally, to know the people's awareness about the Indian government policies, social influence, promotions, web contents plays a significant contribution since people use a very different style of writing the similar contents, an accurate context and drift detection system is needed. The drift in the data makes sample data distribution unbalance, which causes a change in the label either gradually or suddenly.

Thus, preprocessing and identifying more contributing features plays an important role to correctly identify the class label, identifying more relevant features further improves the prediction model, which helps in identifying the content of social media for popularity prediction of various government schemes. In the tweets, it is found that people's perspective is changed with respect to time. Based on it, which policy is more popular is predicted. In future, the number of tweets we may exceed for better results. At present limited no of tweets over the time period are considered. Data collected is only in English whereas in India people may post tweets in many regional languages also. A better NLP and Google APIs may be integrated for labelling the tweets.

In the end, a policy, schemes provided by the government is for the people of India. The policies launched by the government are for the people and if these policies are not reaching out or not making an impact then it might recommend to the government about strategies to re-implement. This topic was deliberately debated and discussed over the internet and social media. Using that data, we can analyse the popularity and need of various such schemes for the citizen of the country. However, the Indian government is taking major steps in social media analysis and awareness about it among the people.

Nomenclatures

$A_i(t)$	Accuracy of classifier at time t
C	Classifier
E	Error rate
$F(x)$	Probability of x instance
P_w	Parameter weight
P_v	Parameter values

Abbreviations

AUE	Accuracy Updated Ensemble
MKI	Make in India
NLTK	Natural Language Tool Kit
SWC	Swachha Bharat (Clean India)

References

1. Data.gov.in. (2016). Open government data (OGD) platform India. Retrieved June, 2016, from https://data.gov.in.
2. Government of India. (2016). Performance dashboard. Retrieved June, 2016, from https://www.mygov.in/.

3. Mausud, M.; Gao, J.; Khan, L.; Han, J.; and Thuraisingham, B.M. (2011). Classification and novel class detection in concept-drifting data streams under time constraints. *IEEE Transaction on Knowledge and Data Engineering*, 23(6), 859-874.
4. Lee, C.C.; Tsai, J.; and Hsieh, C.H. (2008). Detecting drifting concepts on the internet. *Journal of Internet Technology*, 9(3), 229-236.
5. Brzezinski, D.; and Stefanoski, J. (2014). Reacting to different types of concept drift: The accuracy updated ensemble algorithm. IEEE *Transaction on Neural Network and Learning Systems*, 25(1), 81-94.
6. Minku, L.L.; and Yao, X. (2012). DDD: A new ensemble approach for dealing with concept drift. *IEEE Transaction on Knowledge and Data Engineering*, 24(4), 619-633.
7. Masud, M.M.; Chen, Q.; Khan, L.; Aggarwal, C.C.; Gao, J.; Han, J.; Srivastava, A.; and Oza, N.C. (2013). Classification and adaptive novel class detection of feature-evolving data streams. *IEEE Transactions on Knowledge and Data Engineering*, 25(7), 1484-1497.
8. Zhao, X.; Heuvel, W.J.; and Ye, X. (2014). A framework for multi-faceted analytics of user behaviors in social networks. *Journal of Internet Technology*, 15(6), 985-994.
9. Ke, J.; Dong, H.-B.; Liang, Y.-W.; Ai, Y.; and Tan, C.-Y. (2016). STD: An improved social recommendation model with temporal dynamics of social relationships. *Journal of Internet Technology*, 17(5), 863-868.
10. Zliobaite, I.; Bifet, A.; PFahringer, B. and Holmes, G. (2014). Active learning with drifting streaming data. *IEEE Transactions on Neural Networks and Learning Systems*, 25(1), 27-39.
11. Escovedo, T.; Koshiyama, A.; da Kruz, A.A.; and Vellasco, M. (2018). DetectA: Abrupt concept drift detection in non-stationary environments. *Applied Soft Computing*, 62, 119-133.
12. Littlestone, N.; and Warmuth, M.K. (1994). The weighted majority algorithm. *Information and Computation*, 108(2), 212-261.
13. Amarnath, B.; Balamurugan, S.A.A. (2016). Review on feature selection techniques and its impact for effective data classification using UCI machine learning repository dataset. *Journal of Engineering Science and Technology (JESTEC)*, 11(11), 1639-1646.
14. Street, W.N.; and Kim, Y.S. (2001). A streaming ensemble algorithm (SEA) for large-scale classification. *Proceedings of the 7th ACM SIGKDD International Conference Knowledge Discovery Data Mining.* San Francisco, California, United States of America, 377-382.
15. Masud, M.M.; Gao, J.; Khan, L.; Han, J.; and Thuraisingham, B. (2009). Integrating novel class detection with classification for concept-drifting data streams. *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Database*. Bled, Slovenia, 79-94.
16. Oza, N.C.; and Russell, S.J. (2001). Experimental comparisons of online and batch versions of bagging and boosting. *Proceedings of the 7th ACM SIGKDD International Conference Knowledge Discovery Data Mining.* San Francisco, California, United States of America, 176-185.

17. Parker, B.; Mustafa, A.M.; and Khan, L. (2012). Novel class detection and feature via a tiered ensemble approach for stream mining. *Proceedings of the IEEE 24th International Conference on Tools with Artificial Intelligence.* Athens, Greece, 1171-1178.
18. Senaratne, H.; Broring, A.; Schreck, T.; and Lehle, D. (2014). Moving on twitter: Using episodic hotspot and drift analysis to detect and characterize spatial trajectories. *Proceedings of the ACM SIGSPATIAL International Workshop on Location-Based Social Networks.* Dallas, Texas, United States of America, 23-30.
19. Masud, M.M.; Chen, Q.; Khan, L. Aggarwal, C.; Gao, J.; Han, J.; and Thuraisingham, B. (2010). Addressing concept-evolution in concept drifting data streams. *Proceedings of the IEEE International Conference on Data Mining*. Sydney, Australia, 929-934.
20. Singh, T.; and Kumari, M. (2016). Role of text pre-processing in twitter sentiment analysis. *Procedia Computer Science*, 89, 549-554.
21. Lifna, C.S.; and Vijayalakshmi, M. (2015). Identifying concept-drift in twitter streams. *Procedia Computer Science*, 86-94.