# Swarm Optimized Opinion Classification Model for Policy Assessment

**Abhilasha Sharma, Paridhi Sachdeva, Nikhil Arora**

***Abstract***: *A government policy is a scheme launched by the governing body of a nation for the welfare of a particular section of the society or the entire public in general. The impact of such a policy can hence only be determined by the response from its target group. The evaluation of these schemes is often challenging, due to the inability of the government body or organization to collect unfiltered and unbiased feedback from the entire population. The aforementioned task may require a large amount of effort, considerable time and in-depth knowledge of advanced technology. However, with the advent of the information era, it is possible to analyze the sentiments of the public using negligible resources. The internet is rich in freely available unused and unstructured data that can be exploited efficiently for various purposes. One such application is opinion mining which allows the user to extract data from social media websites and categorize it into pre-defined classes. This paper is an attempt to assess one of the most important and current government initiatives- "Digital India", through public sentiments. Digital India is a program launched by the Prime Minister of India to transform the country into a technologically advanced and digitally connected nation. This research work corroborates the use of swarm intelligence or nature-inspired algorithms for feature subset selection during opinion mining, as it results in a substantial reduction in the number of features (and consequently a lesser computation time for model training) and increase in the classification accuracy of the model. Therefore, the aim of this study is to analyze public opinion on "Digital India" campaign to ascertain the success (or failure) of the mission, while at the same time, determine the most suited model for automated evaluation of any government policy in the future.*

***Keywords*** *: Digital India, feature subset selection, opinion mining, swarm intelligence, Twitter.*

## I. INTRODUCTION

The 21st century, also popularly known as the Information Age, is continuously leading to the generation of uncountable quantities of data from millions of sources every day. This data (also referred to as big data) includes market trends obtained from online and offline stores and shopping centers, patient records from hospitals, search patterns of search engine users, temperature variations as recorded by thermometers in various places, traffic details from GPS devices, posts and comments on social media websites, heart rate periodically calculated by body sensors, series of images captured by surveillance cameras, and much more. According to recent statistics [1], internet users alone generate approximately 2.5 quintillion bytes ($10^{18}$ bytes or 1 billion GB) of data every single data. Majority of this data lies unused in the open, losing its worth and potential. If even 5% of this data is appropriately exploited, it can result in a number of advantages such as understanding and targeting customers, improving healthcare, optimizing machine performance, refining law enforcement and so on. [2]

Evaluation of government schemes is another important application of big data to ensure that the public benefits from them. A government scheme is an initiative by the governing body of the nation, launched for the welfare of its citizens, which either benefits them directly, such as the *Saubhagya scheme* for providing electricity to the households, or is taken up to improve the state of the nation in general, such as the *Swachh Bharat Mission* to encourage a cleaner country. For a scheme which directly impacts the public, the effect must be measured in terms of how the public perceives the particular scheme and has been affected by it. A careful analysis of the opinion and accounts of the public is required for the same. A positive response from the citizens indicates the success of the scheme.

Digital India is one such program, started on the 1st of July, 2015 by Prime Minister Narendra Modi, with a goal of providing remote regions with high speed internet connections and refining digital literacy [3]. The aim is to transform the country into a digitally empowered nation in the technological domain. The Digital India campaign has resulted in some significant changes. It has led to increased attendance in firms due to Aadhaar and biometric attendance; it has made cashless money transfer easier with the provision of UPI payments and e-wallets; online booking of passport appointment is now possible through the Passport Seva app; DigiLocker facility under this scheme has also made it easier to carry e-documents anywhere and everywhere. In spite of the various success stories of the initiative, it has still been seen with some skepticism among the public due to a number of reasons. Many fake payment gateways and applications have emerged that dupe the users into sharing their bank account details. Furthermore, according to a recent study, only 24% of Indians own a smartphone [4]. Hence, the rest of the population is unable to

Revised Manuscript Received on April 25, 2020.
* Correspondence Author

**Dr. Abhilasha Sharma***, Department of Computer Science & Engineering, Delhi Technological University, Delhi, India. Email: abhi16.sharma@gmail.com

**Paridhi Sachdeva,** Department of Computer Science & Engineering, Delhi Technological University, Delhi, India. Email: paridhisachdeva98@gmail.com

**Nikhil Arora,** Department of Computer Science & Engineering, Delhi Technological University, Delhi, India. Email: nikhilarora986862@gmail.com

*Retrieval Number: D7892049420/2020©BEIESP*
*DOI: 10.35940/ijeat.D7892.049420*
*Journal Website: www.ijeat.org*

*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication*
*© Copyright: All rights reserved.*

make use of most of the services under Digital India, for eg, mAadhaar, BHIM, onlineRTI, Passport Seva etc. Even the 24% of the population that does own smartphones seems to face a lot of issues with the internet speed and is unable to make use of the available facilities.

India ranks 74th in the world in terms of cellular internet download speeds [5]. So despite the fact that the government has put in a lot of efforts in the development of resourceful applications, not much has been done to ensure that these services reach all the citizens. Lack of digital literacy is another problem in this regard. Additionally, the idea of digitalizing all services also paves way for cybercrimes. Therefore, while there can be arguments both in favor of and against the program, it is important to carefully analyze the public sentiments, so as to determine if the positives outweigh the negatives of the program or vice versa. Hence, the objective of this paper is to recommend, implement and validate an automated feedback mechanism for governance.

The data source chosen for this study is *Twitter* due to a variety of reasons. *Twitter* is a global social networking platform, where people from all regions, religions, castes, creeds, genders, economical backgrounds and professional backgrounds share their views and opinions publicly. Moreover, all the data published by the users is openly visible to everybody across the internet, which makes it easier to get an unbiased data set. Also, it is an extremely fast platform, which means that people’s reactions and thoughts are instantaneously posted and made available online [6]. All these reasons have made it the best choice for selecting *Twitter* as the source of corpus collection for this study.

### A. Proposed Model

The framework for sentiment classification and analysis used in the research is summarized in Fig. 1 and has been described below in detail.

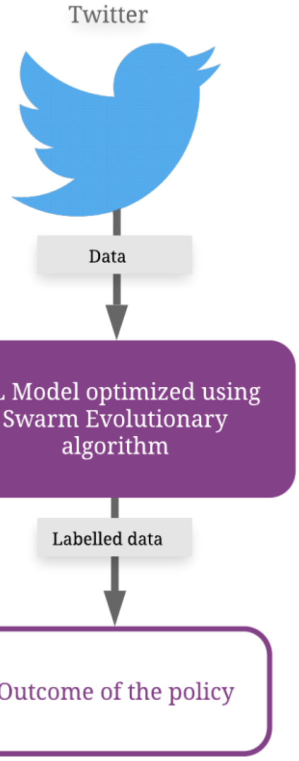


**Fig. 1.Basic framework of proposed model.**

Tweets on the topic are extracted for the required time period from *Twitter*, which form the data set for the problem. This data is split into two sections: the training data and the testing data. After pre-processing this data and extracting the relevant features from it, the data set is input into the Machine Learning (ML) model. The ML model used in this study consists of one of the four machine learning algorithms: Naive Bayes (NB), K-Nearest Neighbors (kNN), Decision Tree (DT) and Support Vector Machine (SVM), further optimized using nature inspired swarm evolutionary algorithms. Swarm algorithms are applied to the existing machine learning algorithms to increase the accuracy of the model through feature subset selection. The aim is to obtain a subset of the features, so as to maximize the accuracy of the model, while reducing the number of features being taken into consideration. For this, accuracy of the model is selected as the fitness function in the swarm algorithm, so that it improves with every iteration of the algorithm. Apart from making the model more accurate, optimizing the model also results in reduced computation time due to a smaller subset of features being used in the model, although it does involve an overhead computational time of obtaining the appropriate feature subset. This optimized model returns categorized data, i.e. data with appropriate labels. The tweets are now categorized as positive, neutral or negative, depending upon their polarity. These statistics can then be studied as a function of time to analyze how the perception of the public has changed with time or with certain developments in the scheme in question. A large number of positive tweets points towards a satisfied public, while a large number of negative tweets indicate that the scheme failed to achieve the impact that was intended.

To increase the accuracy of the model through feature subset selection, only those attributes need to be selected that are significant for the model and give the best result, while dropping the insignificant features. If there are N features, the total possible number of feature subsets is $2^N$. Since finding and testing all these subsets to find the right one is an NP-hard problem, Swarm Evolutionary Algorithms have been used for this purpose. These algorithms have a particular stopping criterion to limit the number of iterations and stop the algorithm from executing for undesirable amounts of time. Swarm algorithms are nature inspired algorithms that study the behavior of groups of organisms in nature and employ the same strategies to computational problems. Two swarm algorithms have been used in this research –Artificial Bee Colony (ABC) and Particle Swarm Optimization (PSO).

### B. Organization

The remaining paper is structured as follows: Section II consists of an overview of related research work and similar sentiment classification models used or proposed in the past by other authors. It also describes the novelty of the framework used in this study. Section III gives the detailed methodology employed in the study with specifications of each step of the program flow. Section IV covers the results and findings of the work. Finally, Section V talks about the conclusion derived from the results of the study, as well as the learnings from the research.

## II. LITERATURE REVIEW

While optimized sentiment classification of data with feature subset selection using swarm evolutionary algorithms has become an upcoming technological tool, this hybrid classification model has been applied to the government intelligence sector for the first time in this paper. This section covers the latest related work carried out by various other authors in the area of opinion mining for policy evaluation, which has also been summarized in Table- I. In 2017, A. Kumar and A. Sharma [7] also conducted a detailed literature survey to review the substantial research in opinion mining in the government intelligence sector. The summarized details of all past work in this domain until 2017 can be found in Table 4 of [7].

In 2016, P. Mishra, R. Rajnish and P. Kumar [8] conducted sentiment classification and analysis on Digital India Mission using dictionary matching approach. This method uses a pre-defined dictionary containing words already segregated under negative, positive and neutral labels. The pre-processed features are matched with the words in the dictionary and labeled accordingly. The reliability of this model is as strong as the dictionary. If the dictionary correctly covers all the words appearing in the data set, the model should have a high accuracy. However, if the dictionary has been obtained from an unreliable or unrelated source (which is usually the case), it might not contain all the words in the data, resulting in a lower accuracy model. Our study makes use of Machine Learning algorithms that use an initial training data set for the training of the classification model, on the basis of which the labels are predicted and tested for the testing data. Furthermore, Mishra, P. et al worked on a data set of 500 tweets, while our study uses 3323 tweets on the same topic. A larger data set spanning a bigger time frame ensures greater accuracy of results.

**Table- I: Related work of opinion mining for policy evaluation**

| Author(s) | Year of study | Topic of data set | Data set description | Feature Selection technique used | Sentiment Classification technique used | Performance Evaluation measure used |
|---|---|---|---|---|---|---|
| ***P.Mishra, R. Rajnish, P. Kumar [8]*** | 2016 | Digital India | 500 tweets | - | Knowledge based approach: Dictionary matching | - |
| ***P. Singh, R. S. Sawhney, K. S. Kahlon [9]*** | 2017 | Demonetization of 500 and 100 rupee bank notes | 30,220 tweets | - | API from meaningcloud as MS Excel add-in | - |
| ***A. Kumar, A. Sharma [10]*** | 2018 | Saubhagya Yojna | 1,262 tweets | - | ML approach: NB, SVM, MLP, kNN, DT | Accuracy, Precision, Recall |
| ***P. Singh, R. S. Sawhney, K. S. Kahlon [11]*** | 2018 | GST implementation | 41, 823 tweets | - | ML approach: NB, SVM, kNN, DT | Accuracy, Kappa Statistics, Matthews correlation coefficient, relative absolute error |
| ***A. Sharma, N. Arora, P. Sachdeva [6]*** | 2019 | CMDRF | 1,666 tweets | - | ML approach: NB, SVM, MLP, kNN, DT, RF, LR, kStar, Adaboost, Bagging | Accuracy, Precision, Recall |

P. Singh, R. S. Sawhney and K. S. Kahlon [9] performed a state-wide, as well as nation-wide opnion mining of demonetization of 500 and 1000 rupee bank notes by the government of India. A total of 30,220 tweets were gathered over a two-phase period of 16 days. For the purpose of sentiment classification, a pre-defined API from meaningcloud was used as an add-in to MS Excel for the classification of tweets into Neu (neutral), N (negative), P (positive), N+ (highly negative) and P+ (highly positive). Our research employs machine learning algorithms on training and testing data for classification of tweets into “Positive”, “Negative” and “Neutral” classes.

A. Kumar and A. Sharma [10] performed opinion mining of Twitter data on Saubhagya scheme using five machine learning algorithms, namely: k-Nearest Neighbors, Decision Tree, Naive Bayes, Support Vector Machine, and Multilayer Perceptron (MLP). Our research undertakes the use of nature-inspired meta-heuristics for further optimization of machine learning algorithms to reduce the number of features obtained. In 2018, P. Singh, R. S. Sawhney and K. S. Kahlon [11] implemented a location based sentiment analysis on the impact of the Goods and Services Tax (GST) reform by the Indian government. They used machine learning algorithms from WEKA 3.8 software for sentiment classification, and compared the same on the basis of accuracy, Kappa Statistics, Matthews correlation coefficient and relative absolute error. In our study, the machine learning algorithms have been compared solely on the basis of accuracy, due to accuracy being used as the fitness function in our swarm algorithms.

A. Sharma, N. Arora and P. Sachdeva [6] proposed a machine learning based opinion mining model for the polarity classification of public tweets on Chief Minister's Distress Relief Fund (CMDRF) in response to the Kerala floods of 2018. A total of 1666 tweets were collected across a time frame of one month. Our current study uses PSO and ABC to reduce the features obtained after feature extraction so as to improve the accuracy of the proposed model, while at the same time, reduce the time taken to train the model.

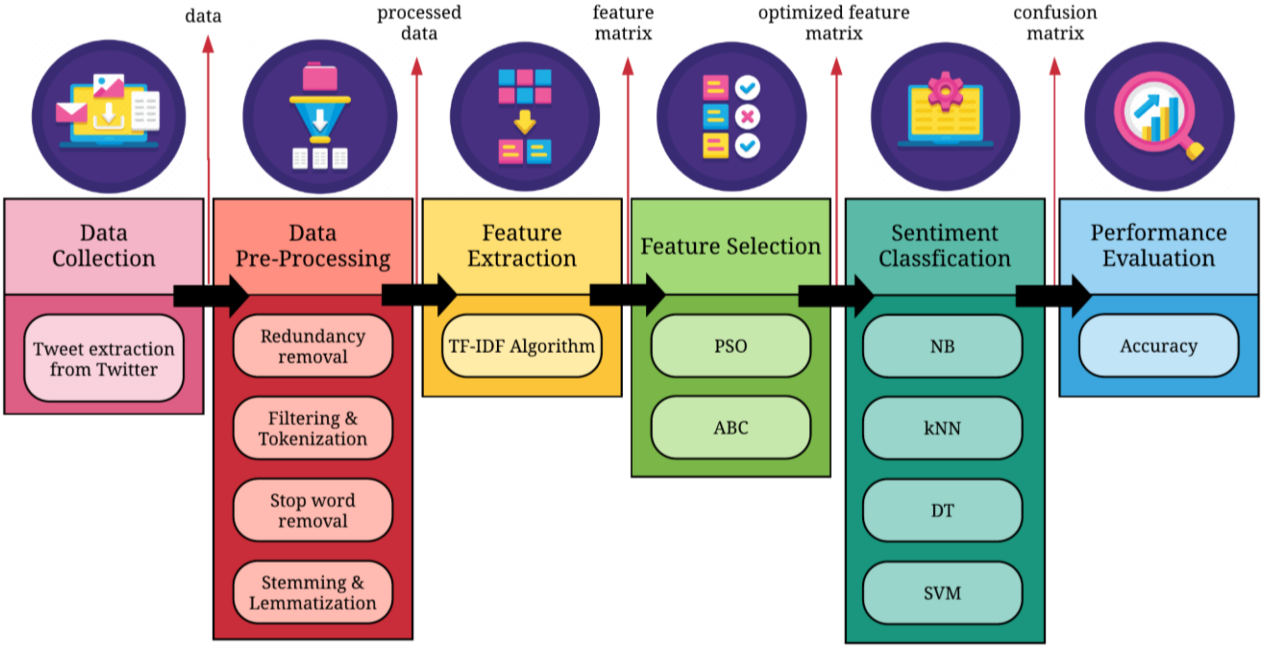


**Fig. 2.System architecture of proposed model.**

## III. METHODOLOGY

The detailed functional flow of the opinion mining model proposed in the research is depicted in Fig. 2. Every step in the process is elaborated in detail in the subsequent sections.

### A. Data Collection

**Table- II: Year-wise distribution of number of tweets collected on "Digital India"**

| Year | Number of tweets collected |
|---|---|
| 2015 | 953 |
| 2016 | 760 |
| 2017 | 452 |
| 2018 | 571 |
| 2019 | 587 |
| ***Total*** | **3323** |

The messages posted by users on *Twitter* are known as "tweets" and can be viewed by anybody globally. Hashtags ("#") are often used in tweets to increase their visibility pertaining to a particular topic, eg. Tweets related to Digital India will commonly be followed by "#DigitalIndia to increase their reach. Tweets containing the hashtag "#DigitalIndia" were extracted using the Tweepy API, after acquiring an authentication key from *Twitter*. These tweets were collected from the first occurrence of #DigitalIndia (i.e. 1st July 2015, which was also the date of launch of this campaign), till the end of 2019 (i.e. 31st December, 2019). A total of 3323 tweets have been collected across this time frame, whose year-wise distribution is given in Table- II and graphically shown in Fig. 3.

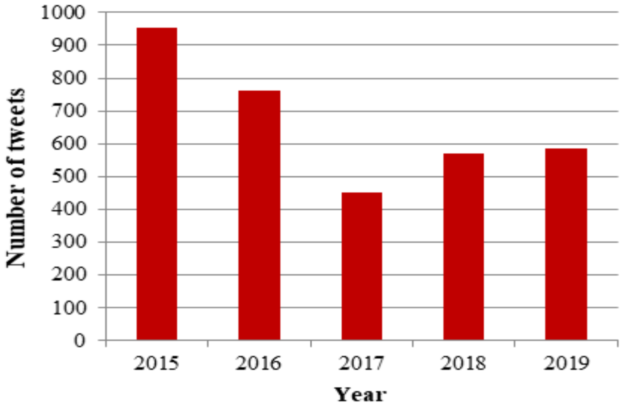


**Fig. 3.Graphical representation of year-wise distribution of number of tweets collected on "Digital India".**

As can be seen from the graph, there was a sudden burst in the number of tweets in the launch year, due to the initiative being of broad and current interest among the public. The trend saw a gradual decline in the next couple of years, followed by a slight increase in the last two years. Due to elections in 2019, people started talking more about the achievements of the government and the topic "Digital India" once again gained momentum.

### B. Data Pre-Processing

For the purpose of representation of data, we use a 'bag of words' approach which contains a list of all the terms appearing in the data set, along with their frequencies.

The data extracted from *Twitter* is raw and unstructured and needs to be cleaned for conversion into a usable and efficient format. This process of removing noise and redundancy from the data and converting it into a high quality data set is referred to as data pre-processing. The techniques used for cleaning of data include [10]:

- *Redundancy removal*: Re-tweets and duplicate tweets are removed, as they may hamper the results
- *Filtering*: Special symbols (@, !, $, * etc.) and URLs are removed as they do not add any meaning to the data set.
- *Tokenization*: Tweets are segmented into a bag of words (separate words) by removing spaces and omitting punctuation marks.
- *Stop word removal*: Stop words are the common but un-informative words in the data set such as "a", "and", "the" etc. These are filtered out.
- *Stemming*: Complex words are reduced to their stems by removing common prefixes, eg. Playing -> Play.
- *Lemmatization*: Complex words are reduced to their root forms by ensuring that the lemmatized word is a dictionary word, eg. Sat -> Sit.

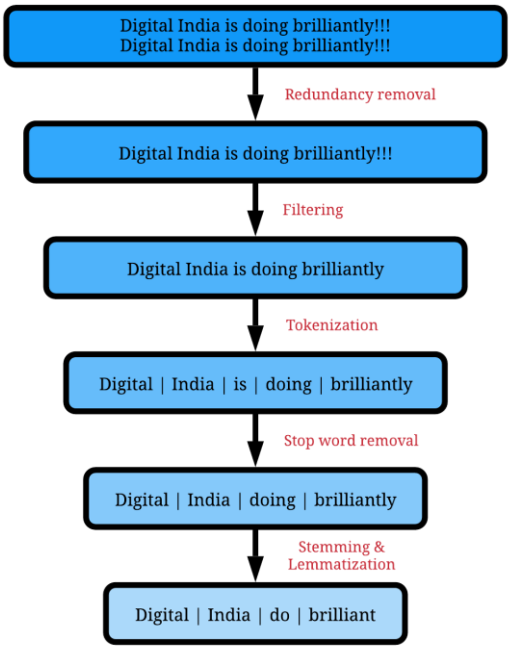

Digital India is doing brilliantly!!!
Digital India is doing brilliantly!!!
Redundancy removal
Digital India is doing brilliantly!!!
Filtering
Digital India is doing brilliantly
Tokenization
Digital | India | is | doing | brilliantly
Stop word removal
Digital | India | doing | brilliantly
Stemming &
Lemmatization
Digital | India | do | brilliant

**Fig. 4.Example of step-wise results obtained from data pre-processing on sample text.**

### C. Feature Extraction

The filtering technique used for extracting features is Term Frequency – Inverse Document Frequency (TF-IDF). TF-IDF algorithm calculates values or weights for each term in a group of documents. A high value of the TF-IDF of a term implies a strong relationship of the term with the document it appears in, suggesting that if that word were to appear in a query, the document could be of interest to the user [12].

*Term Frequency (TF)*: TF refers to the number of occurrences of a term in a particular document, and hence is calculated for a term *t*, with respect to a document *d*. To prevent the value from being dependent on the size of the document, this value is normalized by dividing the actual frequency of the term by the total count of terms in the entire document. The formula for calculation of TF is given in (1).

$$TF(t,d) = \frac{frequency\ of\ t\ in\ d}{total\ number\ of\ terms\ in\ d} \qquad (1)$$

*Inverse Document Frequency (IDF)*: IDF refers to the inverse of document frequency, calculated with respect to a term *t*. IDF is calculated as given in (2). The idea of calculating IDF is to account for those words that provide little information and are yet extremely frequent, such as "is". IDF tells us how concentrated a term is in a particular document. Since the total number of documents may be huge, we take log of this value to prevent explosion [13].

$$IDF(t) = \log \frac{total\ number\ of\ documents}{number\ of\ documents\ that\ contain\ t} \qquad (2)$$

*Term Frequency – Inverse Document Frequency (TF-IDF):* TF-IDF is the multiplicative value of the two terms described above. This value gives the weight associated with a term *t* that describes its importance relative to a document *d* and a collection of all such documents. Using (1) and (2), this value is calculated as given below:

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \qquad (3)$$

### D. Feature Subset Selection

The process of feature subset selection is optional and the results can be obtained with or without this step. The goal of this research is to study the effects of adding this step to determine whether the overhead involved in this phase is worth the efforts and time involved, given the results. The intention is to lessen the number of features in the feature matrix obtained from the previous step, so as to reduce the time taken to train the model. The problem lies in selecting that subset of the features that increases the classification accuracy of the model at the same time. For this purpose, we have undertaken Swarm Evolutionary (SE) algorithms, inspired by organisms in nature and their common behavior. Due to their wide acceptability and varied application area, we have used PSO and ABC for feature subset selection here.

- *Particle Swarm Optimization:* This algorithm is inspired by a swarm of birds in search of food. Each bird knows how far it is from the food source and moves in random directions to reach the food source at the earliest.
  Each bird also keeps track of its personal best location (closest it has been to the food particle) and the global best location of the swarm (closest any bird has been to the food source). Every next step that the bird takes is influenced by these three factors- its current velocity direction, its personal best location, and the swarm's global best location. In this way, the birds collectively reach the food source in the shortest possible time. The PSO algorithm uses the same method of storing and continuously updating two parameters: personal best (pBest) and global best (gBest) to reach the global maxima [14].

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

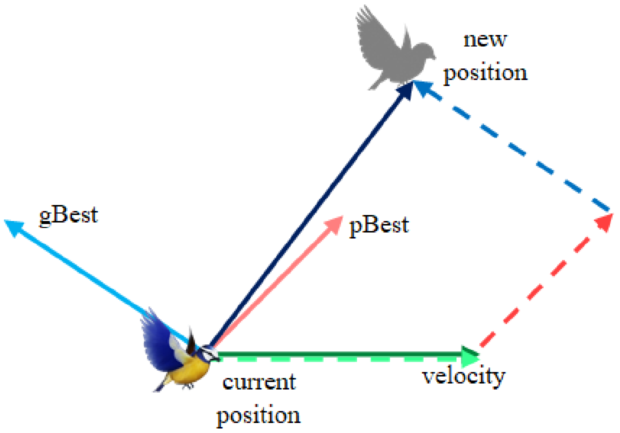


**Fig. 5.Determination of new position of particle using PSO.**

- *Artificial Bee Colony:* This algorithm is inspired by the strategy adopted by a swarm of bees in search of nectar. The bees divide themselves among three functional groups: the scout bees, the onlooker bees and the employed bees. The employed bees are linked with certain food sources, while the scout bees are responsible for randomly looking for a new source and passing the information regarding the quality and usability of the source to the onlooker bees in the dancing area through a waggle dance. The length of the waggle dance determines the quality of the food particle. On the basis of this information, the onlooker bees decide which food source to exploit next [15]. The ABC algorithm uses a similar tactic to obtain the best solution (best food source) for the problem, given a particular fitness function (quality and distance of food source).

All SE algorithms work on a number of initially randomized swarm particles that are constantly trying to reach the global optimum. With every iteration of the algorithm, the individual particles try to optimize the value of the fitness function using certain equations. The fitness function is that value which we are trying to optimize through the SE algorithm, which in this case is the classification accuracy of the model. The point at which all the particles finally converge is the most optimized value of the fitness function, also the global maxima. SE algorithms make use of a stopping criterion or a termination condition that limits the number of iterations. The following are some examples of common termination conditions [16]:

- Maximum number of iterations reached
- Maximum permitted error rate crossed
- Maximum (or minimum) value of fitness function reached
- No improvement in fitness function value for some fixed number of last consecutive iterations
- Maximum CPU time reached

We have used the first two stopping criteria for our model. The two algorithms PSO and ABC have been applied individually with every ML algorithm respectively, since accuracy of the model is dependent on the algorithm used. Hence, the optimized feature matrix obtained for each model is different.

### E. Sentiment Classification

Once the tweets are acquired, they are labeled with the help of certain machine learning algorithms. ML refers to the study of statistical models that teach machines to do certain jobs, on the basis of previously acquired experience in the form of trends or patterns. The training : test data split ratio is taken as 70:30 with 10 fold cross validation. This data is input into four well known machine learning algorithms, namely Decision Tree (DT), Naive Bayes (NB), k-Nearest Neighbors (kNN), and Support Vector Machine (SVM) to train and test these classifiers (pre-defined classification labels used are “positive” or “+1”, “negative” or “-1” and “neutral” or “0”) by segregating the tweets on the basis of their polarity.

- *Naive Bayes:* This is a probabilistic method which calculates the probability of every possible label being the correct label. The one with the highest probability is chosen as the correct label for the given class instance or data.
- *kNN:* In this method, all the data points are hypothetically plotted on a Cartesian plane. Only the closest k points from the given class instance are taken into consideration. This distance is calculated using Euclidean formula, i.e. the shortest straight-line distance between two points. Finally, the label that the majority of these k particles have is given to the class instance in question.
- *Decision Trees*: This algorithm employs the use of a tree structure for deciding the labels. The distinguishing attributes or features form the internal nodes of the decision tree, the presence or absence of these features are the branches and the labels are the leaf nodes. The traversal starts from the root of the tree, and the leaf node at which the traversal ends become the determined class label for the class instance.
- *Support Vector Machine:* This method of classification uses hyperplanes or decision boundaries to separate various classes. The label for the class instance is chosen on the basis of the region in which the instance lies, as defined by the maximal margin hyperplanes surrounding the region.

This phase involves the implementation, analysis and comparison of these four ML models, running on the same data set and (optimized) feature matrix. The goal is to find the best suited supervised learning model for future evaluation of a government scheme.

### F. Performance Evaluation

Once a supervised learning algorithm finishes running, a confusion matrix (or error matrix) is obtained. This matrix (as shown in Table- III) contains the number of correct and incorrect matches for each label respectively and can be used to calculate a number of important performance evaluation measures [17].

**Table- III: Basic format of a Confusion Matrix**

| | | Actual Class | |
|---|---|---|---|
| | | ***Positive*** | ***Negative*** |
| **Predicted Class** | ***Positive*** | True Positive (TP) | False Positive (FP) |
| | ***Negative*** | False Negative (FN) | True Negative (TN) |

Four widely used standard evaluation measures to study performance of ML algorithms include: Accuracy, Precision, Recall and Specificity. Higher values of these measures indicate a better trained model. Accuracy refers to the closeness of the predicted labels with the actual labels [6], and is found to be the most important measure among the above mentioned parameters; hence we have based our study on the calculation, optimization and comparison of accuracy of different classification models. It is calculated from the confusion matrix as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

## IV. RESULTS AND FINDINGS

### A. Opinion Mining of Digital India

Opinion mining is the extraction and study of the beliefs, views, thoughts, sentiments and perceptions of the people. It aims at extracting, collecting, analyzing and classifying all these opinions, on the basis of the underlying sentiments. A similar process has been carried out in this study. A total of 3323 tweets were collected across 5 years on "#DigitalIndia" and were classified into three categories- "positive" or "+1", "negative" or "-1" and "neutral" or "0", on the basis of the polarity of the messages they carry. Table- IV shows the results of opinion mining on this data.

**Table- IV: Polarity of tweets across entire time frame**

| Sentiment | Number of samples |
|---|---|
| Positive (+1) | 1549 |
| Negative (-1) | 816 |
| Neutral (0) | 958 |
| ***Total*** | **3323** |

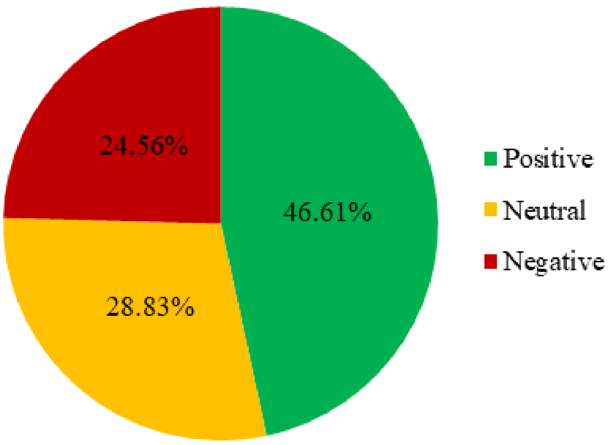


**Fig. 6.Graphical representation of polarity of tweets across entire time frame.**

Fig. 6 represents the complete distribution of the total number of tweets across the entire time period. 46.6% of the total tweets were found to be in support of or encouraging the Digital India campaign, while 24.5% of the public was against the policy or dissatisfied with its implementation. 28.8% of the tweets were found to be merely informational, not portraying any kind of sentiment. Table- V gives a year-wise analysis of the polarity of the tweets.

**Table- V: Year-wise polarity of tweets**

| Year | Sentiment (number of samples) | | | *Total* |
|---|---|---|---|---|
| | *Positive (+1)* | *Negative (-1)* | *Neutral (0)* | |
| 2015 | 411 | 264 | 278 | **953** |
| 2016 | 393 | 205 | 162 | **760** |
| 2017 | 202 | 144 | 106 | **452** |
| 2018 | 287 | 170 | 114 | **571** |
| 2019 | 256 | 175 | 156 | **587** |
| ***Total*** | **1549** | **958** | **816** | **3323** |

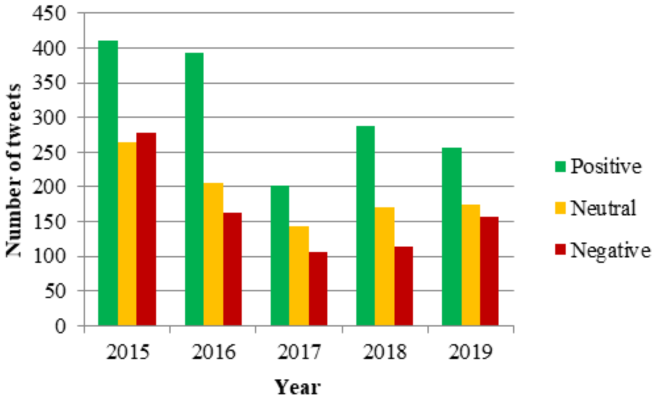


**Fig. 7.Graphical representation of year-wise polarity of tweets.**

**Fig. 8.**

As can be seen from Fig. 7, the number of tweets with positive polarity exceeded the number of neutral and negative tweets in all the years. A slight decline can be seen in the trends of all three labels because when a scheme is first launched, it gains a lot of popularity; however it loses the spotlight as time passes. Usually, there are more neutral tweets in the beginning as informative messages about the initiative are shared by a number of sources; however Digital India is an umbrella scheme that has multiple different policies under it, which were launched at different instances of time throughout this period. Therefore, the neutral tweets are also evenly spread across the time frame. The overall trend of Fig. 6 can be compared with that of the last two years from Fig. 7. This is explained by the elections that took place in 2019, which made people critically analyze the work of the government in the late 2018 and early 2019, by carefully weighing the pros and cons of all the government initiatives, including those of Digital India. In conclusion, during the early years of any policy, the public usually has some preliminary views and thoughts that tend to change over time. The opinions in the latter years majorly represent a broader perspective or analysis of the entire campaign.

## B. Empirical Analysis of different models

Our research was carried out with 12 different combinations of models, in order to find the most suitable one for future assessment of a government scheme. The performance of the tested models has been compared on the basis of two important results:

- Reduction in the number of features after feature subset selection
- Improvement (or deterioration) in the accuracy of the model with feature subset selection

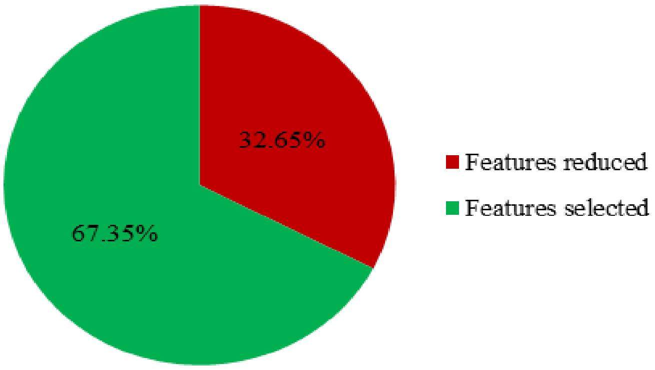
i. Feature Reduction

In the third stage of the flow diagram in Fig. 2, feature extraction was performed by applying tf-idf algorithm on the data set, to obtain a group of features. 886 features were attained in this manner. In the next stage, swarm intelligence algorithms were applied for feature subset selection. The number of feature subsets obtained after applying PSO and ABC respectively, along with the percentage of features reduced has been displayed in Table- VI.

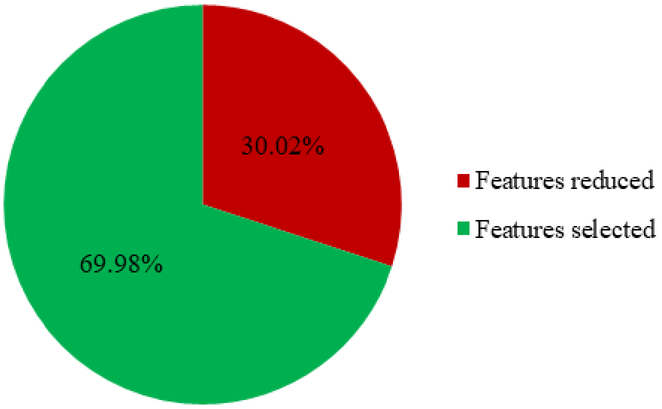
**Table- VI: Number of features selected across different models after feature extraction and feature selection**

| Algorithm | | | NB | kNN | DT | SVM |
|---|---|---|---|---|---|---|
| **Feature Extraction tf-idf (# features)** | | | 886 | 886 | 886 | 886 |
| **Feature Selection** | ***PSO*** | ***Features Selected (Number)*** | 512 | 428 | 704 | 743 |
| | | ***Features Reduced (%)*** | 42.21 | 51.69 | 20.54 | 16.14 |
| | ***ABC*** | ***Features Selected (Number)*** | 472 | 529 | 784 | 695 |
| | | ***Features Reduced (%)*** | 46.73 | 40.29 | 11.51 | 21.56 |

On applying PSO (with tf-idf), the minimum number of features selected was 428 out of 886 for kNN, which is a 51.69% reduction in the number of features. Contrastingly, the maximum number of features selected was 743 for SVM with a 16.14% reduction only. On the other hand, ABC showed the best feature subset selection results on NB with 472 features, which is a 46.73% reduction, while the maximum number of features selected was 784 for DT, with a mere 11.51% reduction. Fig. 8 gives a graphical representation of the average feature subset reduction using PSO and ABC respectively.



(a)



(b)

**Fig. 9.(a)Average feature reduction using PSO. (b)Average feature reduction using ABC.**

As can be seen from the figures above, the average reduction in the number of features after applying tf-idf and PSO was 32.65%, while the average reduction after applying tf-idf and ABC was 30.02%. The percentage of features selected is directly proportional to the time taken to train the model. If the number of features being input into the model is less, the computation time taken by the model will also be less. Hence, the percentage of reduction of features is an important parameter while comparing feature subset selection algorithms. Even though PSO showed better results on the basis of average number of features selected and reduced, this factor alone is insufficient to choose one algorithm over another. Therefore, in the next sub-section, we compare the algorithms and models on the basis of their accuracy.

### ii. Impact on Accuracy

Accuracy is one of the most important performance evaluation measures that is used to compare and contrast the performance of various machine learning models. The following table shows the accuracies of different ML algorithms, with and without the two methods of feature subset selection.

**Table- VII: Accuracy obtained across different models after feature extraction and feature selection**

| Algorithm | | NB | kNN | DT | SVM |
|---|---|---|---|---|---|
| **tf-idf** | ***Accuracy (%)*** | 72.56 | 78.82 | 74.45 | 87.60 |
| **tf-idf + PSO** | ***Accuracy (%)*** | 86.97 | 89.75 | 91.38 | 90.37 |
| | ***Accuracy Improvement*** | 14.41 | 10.93 | 16.93 | 2.77 |
| | ***Accuracy Gain (%)*** | 19.86 | 13.87 | 22.74 | 3.16 |
| **tf-idf + ABC** | ***Accuracy (%)*** | 88.43 | 90.24 | 87.57 | 89.21 |
| | ***Accuracy Improvement*** | 15.87 | 11.42 | 13.12 | 1.61 |
| | ***Accuracy Gain (%)*** | 21.87 | 14.49 | 17.62 | 1.84 |

*Retrieval Number: D7892049420/2020©BEIESP*
*DOI: 10.35940/ijeat.D7892.049420*
*Journal Website: www.ijeat.org*

*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication*

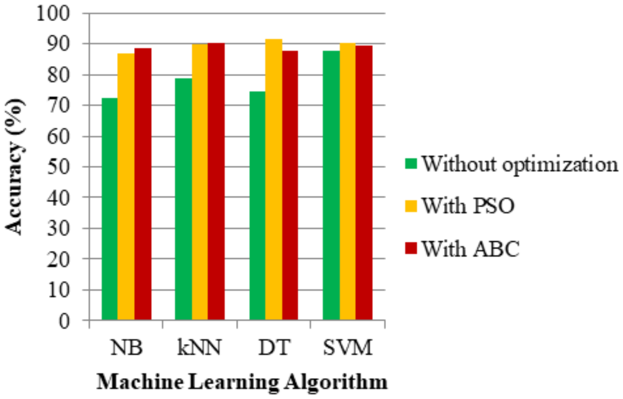


**Fig. 10. Graphical representation of accuracy obtained across different models after feature extraction and feature selection.**

The above graph clearly depicts how the accuracy improved on applying swarm evolutionary algorithms for feature subset selection. As can be seen from Table- VII, before applying feature selection, SVM performed the best with 87.60% accuracy, however, after applying feature subset selection algorithms, DT with tf-idf and PSO showed the highest accuracy of 91.28%. This model also showed the maximum increase in accuracy (16.93%) and highest accuracy gain (22.74%) across all the different combinations. Overall, applying PSO along with tf-idf resulted in an average of 11.26% improvement, while applying ABC with tf-idf showed 10.50% improvement.

## V. CONCLUSION

Digital India is a mission launched on 1st July, 2015 by the Indian government, with a goal to make the country digitally empowered by making all of its services online and reachable to all its citizens. The aim of this paper was to examine the public's perception of this campaign, and at the same time, propose a suitable model for future appraisal of similar policies. Through opinion mining of tweets on "#Digital India", it was found that 46.6% of the people are in favor of the program, while 24.5% of the public is hesitant about the initiative. Thus, while the campaign is still active, this model will not only help the government in understanding its impact on the people, but will also give direction for future decisions and courses of action.

The model for opinion mining proposed in this research involves six major steps, namely (i) Data Collection from *Twitter* (ii) Data pre-processing (iii) Feature Extraction using tf-idf algorithm (iv) Feature Subset Selection using PSO or ABC (v) Sentiment Classification using Machine Learning algorithms (vi) Performance Evaluation of models on the basis of accuracy. A total of 3323 tweets were collected on "#Digital India" from *Twitter*. This data was pre-processed and converted into a bag of words representation. Next, 886 features were extracted from this data set using tf-idf algorithm. For feature subset selection, PSO and ABC were applied and their results were compared. Applying PSO led to a 32.65% reduction in features and ABC gave a reduction of 30.02%. Afterwards, four machine learning algorithms, namely NB, KNN, DT and SVM were applied and their results were compared on the basis of the accuracy of these models. Without feature subset selection, SVM showed the highest accuracy of 87.6%, however, on applying PSO with tf-idf on DT, the accuracy increased from 74.45% (without feature selection) to 91.38%, which was the maximum accuracy shown by any of the models.

Therefore, our study suggests that using Decision Tree classifier, along with tf-idf algorithm for feature extraction and Particle Swarm Optimization for feature subset selection is the most suitable model for the assessment of government schemes on the basis of the opinion of the public. Furthermore, PSO is a better option for feature subset selection over ABC as it depicted more reduction in the number of features (32.65% average reduction) and subsequently a faster alternative for model training, as well as more increase in accuracy (11.26% average increase).

The proposed framework has a wide scope and can be extended and implemented in a number of different application areas such as business analytics, healthcare systems, fault detection etc. At the same time, the current recommendation can also benefit from small variations, whose impacts could be huge. This includes exploring various other swarm intelligence algorithms like cuckoo search algorithm, bacteria foraging algorithm, bat-swarm algorithm, firefly algorithm etc. Apart from nature inspired meta-heuristics, various other feature subset selection techniques are also gaining popularity. These include ontology, fuzzy logic, rough set theory, genetic algorithm and branch and bound algorithm [18]. Hybrid classifiers can also be implemented for the purpose of opinion mining.

## REFERENCES

1. "16 Big Data Statistics - The Information We Generate [2020]," Tech Jury, 18-Feb-2020. [Online]. Available: https://techjury.net/stats-about/big-data-statistics/#gref. [Accessed: 25-Mar-2020].
2. L. V. Satyanarayana, "A Survey on challenges and advantages in big data," *IJCST*, vol. 6, no. 2, pp. 115-119, 2015.
3. S. Dua, "Digital India: opportunities & challenges," *International journal of science technology and management*, vol. 6, no. 3, p. 6, 2017.
4. V. B. Krishnan, "24% of Indians have a smartphone, says Pew study," The Hindu, 08-Feb-2019. [Online]. Available: https://www.thehindu.com/news/national/24-pc-of-indians-have-a-smartphone/article26212864.ece. [Accessed: 25-Mar-2020].
5. "List of countries by Internet connection speeds," Wikipedia, 18-Mar-2020. [Online]. Available: https://en.wikipedia.org/wiki/List_of_countries_by_Internet_connection_speeds. [Accessed: 25-Mar-2020].
6. A. Sharma, N. Arora and P.Sachdeva, "Machine Learning Based Social Big Data Mining for Communal Welfare," *International Journal of Information Systems & Management Science*, vol. 1, no. 1, 2018.
7. A. Kumar and A. Sharma, "Systematic Literature Review on Opinion Mining of Big Data for Government Intelligence," *Webology*, vol. 14, no. 2, 2017.
8. P. Mishra, R. Rajnish and P. Kumar, "Sentiment analysis of Twitter data: Case study on digital India," *2016 International Conference on Information Technology (InCITe) - The Next Generation IT Summit on the Theme - Internet of Things: Connect your Worlds*, Noida, 2016, pp. 148-153.
9. P. Singh, R. S. Sawhney, and K. S. Kahlon, "Sentiment analysis of demonetization of 500 & 1000 rupee banknotes by Indian government," *ICT Express*, vol. 4, no. 3, pp. 124-129, 2018.
10. A. Kumar and A. Sharma, "Opinion Mining of Saubhagya Yojna for Digital India," *International Conference on Innovative Computing and Communications*, Springer, Singapore, 2019, pp. 375-386.

*Retrieval Number: D7892049420/2020©BEIESP*
*DOI: 10.35940/ijeat.D7892.049420*
*Journal Website: www.ijeat.org*

*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication*

11. P. Singh, R. S. Sawhney, and K. S. Kahlon, "Twitter based sentiment analysis of GST implementation by Indian government," in *Digital business*, Springer, Cham, 2019, pp. 409-427.
12. J. Ramos, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, pp. 133-142, 2003.
13. W. Scott, "TF-IDF for Document Ranking from scratch in python on real world dataset," Medium, 21-May-2019. [Online]. Available: https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089. [Accessed: 25-Mar-2020].
14. J. Kennedy and R. Eberhart, "Particle swarm optimization," *Proceedings of ICNN'95 - International Conference on Neural Networks*, Perth, WA, Australia, 1995, vol. 4, pp. 1942-1948.
15. D. Karaboga and B. Basturk, "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm," *Journal of global optimization*, vol. 39, no. 3, pp. 459-471, 2007.
16. N. M. Kwok, Q. P. Ha, D. K. Liu, G. Fang and K. C. Tan, "Efficient particle swarm optimization: a termination condition based on the decision-making approach," *2007 IEEE Congress on Evolutionary Computation*, Singapore, 2007, pp. 3353-3360.
17. "Confusion matrix," Wikipedia, 10-Mar-2020. [Online]. Available: https://en.wikipedia.org/wiki/Confusion_matrix. [Accessed: 25-Mar-2020].
18. J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," in *IEEE Intelligent Systems and their Applications*, vol. 13, no. 2, pp. 44-49, March-April 1998.

## AUTHORS PROFILE

**Dr. Abhilasha Sharma** is currently working as an Assistant Professor in Department of Computer Science & Engineering at Delhi Technological University, Delhi, India. She has 11 years of work experience in industry, research and academics. She has received her M.Tech. (Master of Technology) and B.Tech. (Bachelor of Technology) degrees in Information Technology. She has completed her PhD in Information Technology from Delhi Technological University, Delhi, India. She has many publications to her credit in various journals and international conferences. Her research area includes Web Applications, Web Engineering, Opinion Mining, Social Web, Big Data Analytics, Social Web based Predictive Modeling.

**Paridhi Sachdeva** is pursing B. Tech. in Computer Science and Engineering from Delhi Technological University, Delhi, India. Her research paper titled "Machine Learning based Social Big Data Mining for Communal Welfare" was published in *International Journal of Information Systems & Management Science* in December, 2018. She is also a recipient of the title of "Child Scientist" at the national level of National Children's Science Congress. Her fields of interest include Machine Learning, Malware Detection and Algorithms.

**Nikhil Arora** is pursing B. Tech. in Computer Science and Engineering from Delhi Technological University, Delhi, India. His research paper titled "Machine Learning based Social Big Data Mining for Communal Welfare" was published in *International Journal of Information Systems & Management Science* in December, 2018. He is the co-founder of International Organization of Software Developers, India's leading open source development based organization. His fields of interest include Machine Learning, Web Development and Data Structures.