International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 5 (2017), pp. 781-790

http://www.ripublication.com

A Study on Digital India Programme Using Azure Cloud and Twitter Data

M.Ganeshkumar
M.Phil. Research Scholar, S.C.S.V.M.V. University,
Kanchipuram-631 561, India.

Dr. V. Ramesh
Assistant Professor, S.C.S.V.M.V. University
Kanchipuram-631 561, India.

Abstract

Digital India is a campaign launched by the Government of India with a vision to transform India into a digitally empowered society and knowledge economy. It consists of three core components. These include, the creation of digital infrastructure, delivery services digitally and digital literacy. This plan will really ensure the growth and development in India especially in the rural areas by connecting rural regions and remote villages with high-speed internet services. During the implementation of these projects Government is required to know the opinions, ideas and feedback of the people. The aim of this study is to analyze the Twitter data about implementation of Digital India program with Azure Cloud ML and provide recommendations to the Government. Twitter data is analysed using different algorithms like k-means clustering, regression and classification algorithms. In this study Logical Regression, Locally Deep Support Vector Machine, Two-Class Support Vector Machine and the Two-Class Bayes Point Machine algorithms were applied to the data set and results were compared to choose a suitable algorithm, which will be used to analysis.

Keywords: Twitter Data, Digital India, Text Analysis, Microsoft Azure ML Cloud.

I. INTRODUCTION

Indian Government has launched the 'Digital India' programme to make the Government's services available electronically to the people of India living in every nook and corner of the country. The main focus of this Digital India is to improve the rural internet connectivity throughout India so that people an avail of the services of the Government. This plan will really ensure the growth and development in India especially in the rural areas by connecting rural regions and remote villages with high-speed internet services.

In Belgium, France, and Canada over 90% of consumer payments are made via cashless modes. The United Kingdom, Sweden, Australia, Netherlands, and the US also have high rates of consumer payments (80% and over) made via non-cash modes. Germany and South Korea also use cashless payments as the major mode of consumer payments. The latter is the only Asian country featuring in the list of the top 10 cashless societies while no country from Africa or South America finds a position in the list. Only time will tell if Indian citizens also favor cashless transaction methods over cash transactions and the effects of cashless transactions on the Indian society and economy. According to CLSA 68% of transactions in India are cash-based, Russia and Indonesia are ahead of India on this count.

To make the program success, Government of India needs to know the opinion, ideas and feedback of the people to improve the services. By knowing the feedback the government can predict or modify the way it is being implemented. Now in India the penetration of internet is very vast even people who are living in remote villages have mobile phones and also having smart phones with internet connection. Customer feedback is very important for any business to succeed and to implement any new ideas. Instead of print media electronic media can be used to get the feedback from the users. Social media plays a vital role as they give direct input of the population quickly whereas print media consumes a lot of time. So to get quick feedback from the customer, social media sites are used.

Among the social media, Twitter receives more attention and it is a main tool for doing research due to its influence in any area. Comparing with Facebook and Instagram, where sharing of images and videos can happen along with text, Twitter was chosen since in this study the main focus is on text. Also people use very short, suitable words to share their opinion. This study analysed the impact of the Digital India on the people. Feedback has been received from the people using the Twitter which is essentially raw data, then it had to be cleaned and analysed using R programming in the Machine Learning of Microsoft Azure Cloud and display the outcome.

The main objective of this study is to study reaction of the people on Digital India campaign and to frame recommendations to the government. Further this study aims to find the suitable algorithm to analyse twitter data.

II. METHODOLOGY

Government of India also has Twitter account for the Digital India where it shares it ideas, projects and other details to the public. Public can express their opinion, ideas about them and provide feedback for the Government which can take necessary or suitable action. Twitter also provides instant information dissipation comparing the other media like Newspaper, Television or Radio where we have to pass through lot of channels to get the message to the People, but Twitter spreads the information and also gets instant output from the Public it acts like two way operations. Since, the main aim of this study is to do an analysis of Digital India tweets, the following steps are involved in the methodology.

- Retrieve tweeter raw data
- Text processing using R tool
- Feature engineering
- Analyse the data using Azure machine learning
- Evaluate model performance

Chorus Tweetcatcher Desktop Edition is used to collect tweeter data. Feature hashing module in Tweetcatcher transform a set of English characters to a set of features representing integers. Text can be represented as numeric feature vectors of equal length and dimensionality can be reduced. Since the output is available as numeric it is also relatively easy to use machine learning methods like classification as it replaces the complex string operations with hash lookups. This makes the feature hashing faster. After capturing the data using this tool, convert the data in to excel format with the tweet text and the sentiment which can be positive or negative or neutral. Twitter account should be created to receive the credentials and for data collection.

Use cloud computing

The main advantage of the cloud is that any researcher with internet connection, user name, password and a system can view the final result from anywhere in the world. If needed, anyone can also do the reengineering and test the outcome. The final evaluated result can be made available as a web service to an application which can make use of it. The physical distance between the scholars and the guides are not present.

While doing the analysis of twitter data, the traditional data mining techniques have their own disadvantages like storing information in physical desktop, limitation of hardware, software, accessibility of the system by other users. The desktop system may not able to handle massive amount of data i.e. Big Data. The configuration of the Hardware and software which involves huge investment cannot be increased always. So the alternative available is Cloud Computing. Cloud computing can be used to do the analysis of the data with the application in the cloud. In Cloud environment third party data centers can be used to store and analyze data. It provides high computation power at low cost where scalable and available applications can be created.

Microsoft Azure Cloud Computing for Twitter Analysis

With a hotmail account, Microsoft Azure cloud can be accessed, it also provides Machine language, R programming as a part of the cloud. Performing a Twitter data analysis using Microsoft Azure cloud computing has lot of advantages. We can store raw data and we can use Machine Learning with R to do the processing and mining of knowledge from data. Azure ML is build on the top of many of the Microsoft products and services. It makes the job of the researcher easier and it is available on the cloud to do predictive analysis. Using Metadata Editor Module in Azure Cloud is used to transform the data into uniform data.

Tweets which are unstructured usually required some preprocessing before it can be analysed. In the study 'R' tool is used to remove punctuation marks, special character and digits and then perform case normalization. Within Machine Learning Studio 'R' scripts are run to preprocess the data.

Feature hashing module is used to transform a set of English characters to a set of features representing integers. Feature hashing module initially creates a dictionary of n-grams. After the building the dictionary, the Feature Hashing module convert all dictionary terms into hash values, and computes whether a feature was used in each case. Azure ML Studio module is used to split the data into training set and test set.

Process of Twitter Data Analysis

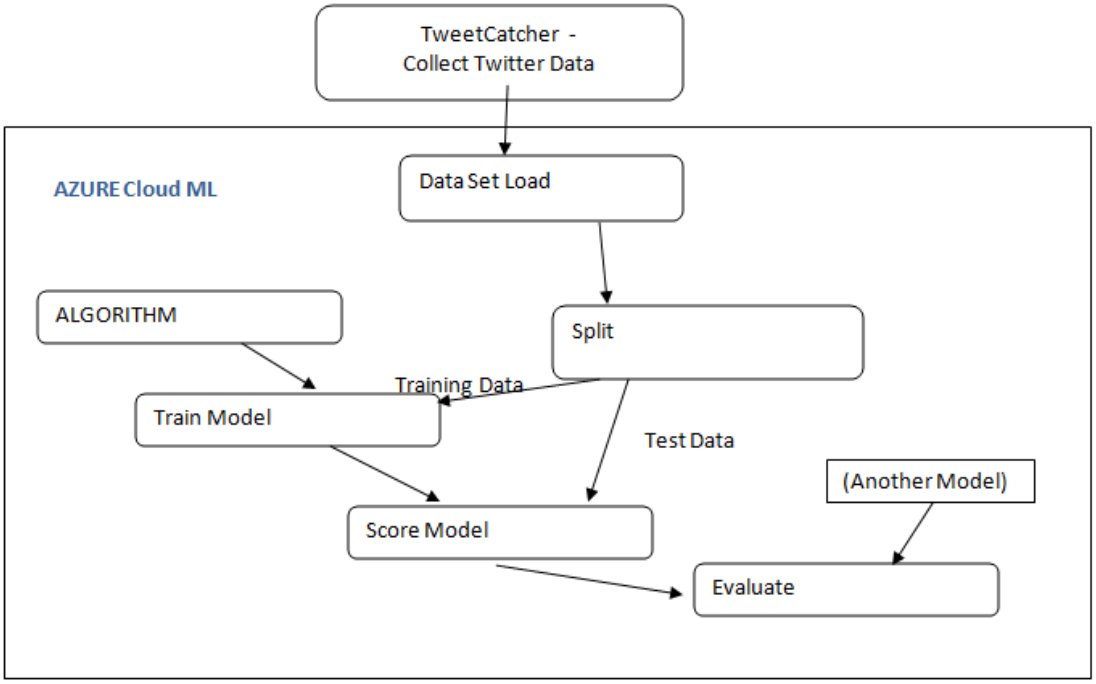

Figure 2.1 Process of Twitter Data Analysis

Comparison of Classification algorithms

In this model for predicting the tweets about Digital India, four classification algorithms are compared and among them the algorithm with maximum accuracy was used. Using the algorithm the feedback of the people was calculated and the prediction was made. This work based on classification can be applied to get the final result of Twitter analysis. Each algorithm gave the corresponding values for the given data set, based on different parameters algorithms are compared and the final result was found.

After comparing the algorithm model it was predicted that the Two class bayes point algorithm has shown most accuracy then using this algorithm an analysis was done, it was found that tweets about Digital India has a overwhelming positive effect on the people.

Metrics used for Comparison

- Accuracy measures the goodness of a classification model as the proportion of true results to total cases.
- Precision is the proportion of true results over all positive results.
- Recall is the fraction of all correct results returned by the model.
- F-score is computed as the weighted average of precision and recall between 0 and 1, where the ideal F-score value is 1.
- AUC measures the area under the curve plotted with true positives on the y axis and false positives on the x axis. This metric is useful because it provides a single number that lets you compare models of different types.

Metrics obtained using different Algorithms

Two Class Support Vector Machine

The threshold value was 0.5 and AUC was about 0.766 which was greater then AUC which is given in the above image. During the evaluation tweets it was found True Positive value is 2100 and the False positive result was about 435 False Negative was 44 and True Negative was 15 as given in the figure 4.4 and also in the below table.

Table 2.1: Two Class Support Vector Machine

True Positive	False Positive	True Negative	False Negative
2100	435	15	44

Two Class Bayes Point Machine Result

In the figure 4.5 for the threshold value of 0.5 was obtained a AUC with 0.990 value with the Two Class Bayes Point Machine algorithm.

True Positive	False Positive	True Negative	False Negative
2144	30	420	0

During the evaluation of tweets with Two Class Bayes Point Machine it was found that value for the True Positive was found with 2144 and for the False positive result the value was about 30 , and True Negative value was 420. For the False Negative the value was zero which was given in the above table.

Two Class Locally Deep Support Vector Machine

True Positive	False Positive	True Negative	False Negative
2144	12	438	0

For the same tweets when the **Locally Deep Support Vector Machine** method was applied again the True Positive was 2144 and False Positive was about 12. True Negative was 438 and False Negative was 0. The Threshold value was 0.5 and the AUC was about 0.997 which was above the Threshold.

Logistic Regression

True Positive	False Positive	True Negative	False Negative
2138	320	130	6

For the same tweets when **Logistic Regression** method was applied again the True Positive was 2138 and False Positive was about 320, True Negative 130 and False Negative 6. The Threshold value was 0.5 and the AUC was about 0.940 which was above the Threshold. It can be concluded since True Positive value was greater than the False Positive, it ensures that the majority of the tweets support the Digital India.

In this study Two-Class Support Vector Machine, Two-Class Bayes Point Machine, Two-Class Locally Deep Support Vector Machine are compared with the above metrics.

Result for the Metrics and Feedback Comparison Algorithms

Metrics	High Value	Least Value
Accuracy	Bayes Point Machine	Support Vector Machine
Precision	Locally Deep SVM	Support Vector Machine
Recall	Support Vector Machine	Locally Deep SVM
F1-Score	Bayes Point Machine	Support Vector Machine
AUC	Bayes Point Machine	Support Vector Machine

- For High Accuracy, F1-Score and AUC Bayes Point Machine was suitable and SVM came second.
- If needed more Precision then Locally Deep SVM can be chosen, here Bayes Point Machine came second position.
- For Recall with higher value Support Vector Machine came first. here Bayes Point Machine took second position.

Result Feedback - Comparison Algorithms

Algorithm	True Positive	False Positive	True Negative	False Negative
Two Class Support Vector Machine	2100	435	15	44
Two Class Bayes Point Machine	2144	30	420	0
Two Class Locally Deep Support Vector Machine	2144	12	438	0
Two Class Logistic Regression	2138	320	130	6

Here from the above table it was found that the Two Class Bayes Point and Two Class Locally Deep Support Vector Machines gave more number of True Positives. Two Class Support Vector Machine predicted least number of True Positive.

CONCLUSION AND RECOMMENDATIONS

Digital India is the initiative started by the Indian government which will implement the services of the government digitally and it is going to transform the way services are being used by the people. Digital India's successful implementation will pave way

for the economic advancement of poor people , as it provide new avenues for the citizens of the country, it is going to be a Digital Revolution which will impact the lives of all the people of India positively. As per the study carried out the total number of feedback for positive, negative and neutral are given in the below figure.

In the below table total positive, neutral and negative values received using the Tweet Catcher are displayed.

Table Result - Tweet Catcher

	Positive	**Negative**	**Neutral**
Counts	751	450	1393
Percentage	28.95	17.35	53.70

Using the above data , in this study different algorithms are compared with each other and Two Class Bayes Point Machine found to be suitable. Using Two Class Bayes Point Machine algorithm it was found that there 2144 True positives which was about 82.65%.

Table Result -Bayes Point Machine

Algorithm	**TP**	**FP**	**TN**	**FN**
Two Class Bayes Point Machine	2144	30	420	0
Percentage	82.65	1.16	16.20	0

In this study the value of True Positive is equal or more than the Threshold value as it is an indicator, that majority of the people who tweeted their opinion have given positive feedback about the Digital India. This result is an indicator of success of the project implementation in the perception of the target population.

SCOPE FOR FURTHER STUDY

Twitter is relatively popular in ease of collecting data about Digital India, but there are limitation as its implementation is covering remote villages, getting feedback from the people in those villages where Digital India has not reached is difficult. Even if it is reached people may not able to give feed back if they are not using any electronic devices. During the implementation in villages getting tweets may be difficult as people may not be interested in interacting initially. So, alternative way of data collected may be considered.

In India there are many languages being used in the mobile apps devices. If the beneficiary of the Digital India is using some other social media sites then getting their inputs about the initiative is challenging. Further research can be carried by making the experiment model as a web service and it can be used by the applications.

REFERENCES

1. Varsha Sahayak , Vijaya Shete , Apashabi Pathan "Sentiment Analysis on Twitter Data" *International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Issue 1, Volume 2 (January 2015)*
2. Analysis Hana Anber1*, Akram Salah2, A. A. Abd El-Aziz31 "A Literature Review on Twitter Data" *10.17706/ijcee.2016.8.3.241-249*
3. Dr.Manoj Kumar Bisht "A New Era Digital India International Journal of *Application or Innovation in Engineering & Management (IJAIEM) Volume 5, Issue 3, March 2016 ISSN 2319 – 4847*
4. Reshma Bhonde1, Binita Bhagwat2, Sayali Ingulkar3, Apeksha Pande "Sentiment Analysis Based on Dictionary Approach" *International Journal of Emerging Engineering Research and Technology Volume 3, Issue 1, January 2015, PP 51-55 ISSN 2349-4395 (Print) & ISSN 2349-4409.*
5. Mining Pushpa Ravikumar, PhD , Adarsh M J "Survey: Twitter data Analysis using Opinion" *International Journal of Computer Applications (0975 – 8887) Volume 128 – No.5, October 2015*
6. Jyoti Siwach & Dr. Amit Kumar "Vision of Digital India: Dreams comes True "*IOSR Journal of Economics and Finance (IOSR-JEF) e-ISSN: 2321-5933, p-ISSN: 2321-5925.Volume 6, Issue 4. Ver. I (Jul. - Aug. 2015), PP 66-71*
7. Pratiksha P. Nikam1, Ranjeetsingh S. Suryawanshi "Microsoft Windows Azure: Developing Applications for Highly Available Storage of Cloud Service " *International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Volume 4 Issue 12, December 2015*
8. Inderjit Kaur, Deep Mann "Data Mining in Cloud Computing " *International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 3, March 2014 ISSN: 2277 128X*
9. Jinal Jani, Girish Tere "Digital India: A need of Hours " *International Journal of Advanced Research in Computer Science and Software Engineering. Volume 5, Issue 8, August 2015 ISSN: 2277 128X*
10. Ravi Sankar G Rajasekhar Reddy V Arun Babu P "Windows Azure: A Highly Available Storage of Cloud Service through Secured Channels" *IJARCSSE Volume 4, Issue 9, September 2014 ISSN: 2277 128X*
11. Data James Spencer and Gulden "Sentimentor: Sentiment Analysis of Twitter " *Uchyigit School of Computing, Engineering and Mathematics University of Brighton, Brighton, BN2 4GJ {j.spencer1,g.uchyigit}@brighton.ac.uk*
12. Digital India: A Vision Towards Digitally Empowered Knowledge economy Indian Journal of Applied Research Volume : 5 | Issue : 10 | October 2015 | ISSN - 2249-555X

13. Michael Collier Robin Shahan *" Microsoft Virtual Academy book Microsoft Azure Essentials "*
14. Jeff Barnes *" Microsoft Virtual Academy book Azure Machine Learning*
15. http://www.digitalindia.gov.in/content/broadband-highways
16. https://en.wikipedia.org/wiki/Machine_learning
17. https://msdn.microsoft.com/en-us/library/azure/dn905974.aspx
18. http://gim.unmc.edu/dxtests/roc3.htm