Comparative Analysis of Various Text Classification Algorithms on WhatsApp Chat Log

Mrs. Jyoti Sharma, Ms. Neelam Sharma, Mr. L. P. Bhaiya
(Department of Computer Science & Eng, Bharti College of Eng. & Tech. Chhattisgarh, India.)

Abstract: **Data Analytics has emerged as an important domain in the digital space due to the explosion of tremendous volume of data by various sources such as social media, sensors and business organizations. Social media contribute in generation of huge varied data formats in various representations. WhatsApp has attracted large volume of users because of the easy chat conversations. In current scenario WhatsApp is used for small scale business, understanding the context in this chat text is important to identify the insights. WhatsApp data is left behind unnoticed as there exist no standard to represent in conventional machine understandable text. In this paper we are studying the various classification algorithms. Classification is the process of dividing the data to some groups that can act either dependently or independently. Our main aim is to show the comparison of the various classification algorithms like KNN, Naïve Bayes, Decision Tree, Random Forest and Support Vector Machine (SVM) with rapid miner and find out which algorithm will be most suitable for the users. The overall results for the all the algorithms are shown in Table1. From the results it is clear that, the K-NN algorithm is better than other algorithms on both the datasets.**

Index Terms **- Data Analysis, Python Programming, Visualization, WhatsApp.**

I. INTRODUCTION

In the current time, because of varied factors like simple use, essential options, the usage of WhatsApp has accelerated. In 2009, WhatsApp was based by Brian Action and Jon Koum[1]. WhatsApp's user base had enhanced to regarding 210 million active users by February 2013, because of the user's ability to move with others through audio vocation, texting, and transferring media likewise as cluster chat [1]. the target of the paper is to classify the amount of users as those confirmed and not smitten by WhatsApp cluster chat and therefore predicting the extent of addiction. above all, this paper in the main emphasizes on the usage of Python applied mathematics computer code software engineer, and the way it will be accustomed extract and work with a selected dataset. Python is Associate in Nursing ASCII text file knowledge analysis setting and programming language [2].It has a good user base in academe specifically and is additionally supported by email and net teams.

1.1. Data Analysis

Process of cleansing, remodeling, inspecting and modeling knowledge with the goal of uncovering helpful data, indicating conclusions, and therefore supporting decision-making is knowledge Analysis [4]. it's multiple facts and approaches encompassing multiple techniques beneath a range of names in disparate business, science, and science domains. Analysis refers to breaking a full part into its separate elements for individual examination. Knowledge analysis may be a method for feat information and reworking it into data helpful for decision-making by users. Knowledge is collected and analyzed for testing hypothesis or responsive the queries. Statistician John Tukey outlined knowledge analysis in 1961 as: "Procedures for analyzing knowledge, techniques for decoding the results of such procedures, ways in which of designing the gathering of knowledge to form its analysis easier, a lot of precise or a lot of correct, and every one the machinery and results of (mathematical) statistics that apply to analyzing knowledge [4]. Phases of knowledge analysis are repetitious and therefore feedback from later phases might lead to extra add earlier phases.

This analysis provides the essential plan of applied mathematics analysis done on a selected WhatsApp cluster knowledge. Following are the sections during which analysis has been carried:

- To find what type of communication medium people prefer the most in WhatsApp group chat.
- To find most active day of week.
- To find which age group participants are more active on WhatsApp group and number of messages send by each age group participants per month, day, hour.
- To find whether Males are more addicted to the WhatsApp group or Females.
- Total number of messages sends as per Timestamp.

II. LITERATURE SURVEY

The dataset of WhatsApp cluster chat used for analysis is of one year (May, 2015 –May, 2016) that consists of 5,563 records in total and includes of bound characteristics that outline what quantity a selected person is victimization WhatsApp Chat cluster, like the years of usage, period of usage in an exceedingly day, the response levels, form of messages denote by every individual within the cluster (Smiley, Text, Multimedia), that age bracket folks are a lot of active then on. the most attributes set for this analysis are form of messages been send, period of use per year/month/week/day /hour, timestamp (AM/PM), age bracket of sender, gender (Male/Female). Jupyter is that the most favored IDE for Python is been accustomed perform exploratory knowledge analysis and visualization for the collected knowledge for the most part owing to its open supply nature.

III. CLASSIFICATION

Classification is a supervised learning technique which places the document according to content. Text classification is largely used in libraries. Text classification or Document categorization has several application such as call center routing, automatic metadata extraction, word sense disambiguation, e-mail forwarding and spam detection, organizing and maintaining large catalogues of Web resources, news articles categorization etc. For text classification many machine learning techniques has been used to evolve rules

(which helps to assign particular document to particular category) automatically [1]. Text classification (or text categorization) is the assignment of natural language documents to predefined categories according to their content. Text classification is the act of dividing a set of input documents into two or more classes where each document can be said to belong to one or multiple classes. Huge growth of information flows and especially the explosive growth of Internet promoted growth of automated text classification [4].

3.1 Classification Methods

3.1.1. Decision Trees Decision tree methods rebuild the manual categorization of the training documents by constructing well-defined true/false queries in the form of a tree structure where the nodes represent questions and the leaves represent the corresponding category of documents. After having created the tree, a new document can easily be categorized by putting it in the root node of the tree and let it run through the query structure until it reaches a certain leaf. The main advantage of decision trees is the fact that the output tree is easy to interpret even for persons who are not familiar with the details of the model [5].

3.1.2. *k-Nearest Neighbor* The categorization itself is usually performed by comparing the category frequencies of the k nearest documents (neighbors). The evaluation of the closeness of documents is done by measuring the angle between the two feature vectors or calculating the Euclidean distance between the vectors. In the latter case the feature vectors have to be normalized to length 1 to take into account that the size of the documents (and, thus, the length of the feature vectors) may differ. A doubtless advantage of the k-nearest neighbor method is its simplicity.

3.1.3. *Bayesian Approaches* There are two groups of Bayesian approaches in document categorization: Naïve [6] and non-naive Bayesian approaches. The naïve part of the former is the assumption of word independence, meaning that the word order is irrelevant and consequently that the presence of one word does not affect the presence or absence of another one. A disadvantage of Bayesian approaches [7] in general is that they can only process binary feature vectors.

3.1.4. *Neural Networks* Neural networks consist of many individual processing units called as neurons connected by links which have weights that allow neurons to activate other neurons. Different neural network approaches have been applied to document categorization problems. While some of them use the simplest form of neural networks, known as perceptions, which consist only of an input and an output layer, others build more sophisticated neural networks with a hidden layer between the two others. The advantage of neural networks is that they can handle noisy or contradictory data very well. The advantage of the high flexibility of neural networks entails the disadvantage of very high computing costs. Another disadvantage is that neural networks are extremely difficult to understand for an average user [4].

IV. SYSTEM ARCHITECTURE

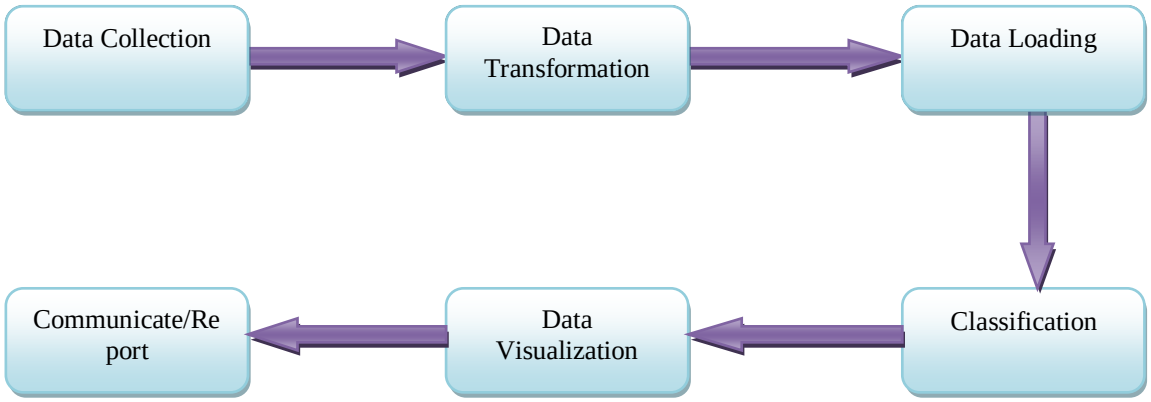

Fig 1: Overview of the System Architecture

Data Analysis process includes Data Collection, Data Transformation, Data Loading, Exploratory Data Analysis, Data Visualization and Communicate/Report. Fig 1 displays the process of Data Analysis.

V. METHODOLOGY

Now let's have a glance on different stages and procedures in retrieval of insightful results, gathering relevant data, importing it into the algorithms and finally analyzing it.

5.1 Data Collection

Data Collection is the first stage of the model which includes idea, defining project objective, setting up machine and lastly knowing your data. Fig 2 illustrates these processes.

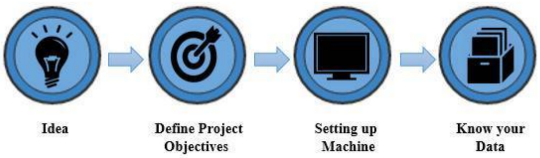

Fig 2: Processes involved in Data Collection

Accurate data collection is essential in order to ensure the integrity of research. The purpose of collecting data is to answer questions in which the answers are not immediately obvious and thus helps in decision making [5].

The person must have absolute idea with regards to the actual source of data, methods of extraction and usefulness of data. In this case the data been extracted is historic data that holds the key to understand data over time. A copy of the history of a group chat is been extracted, using the Email chat feature:

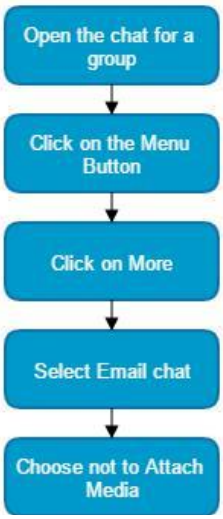

Fig 3: Steps to extract data from WhatsApp using email chat feature 1

A .txt document of your chat history will thus be attached and an email will be composed to the specified sender.

5.2 Data Transformation

Once the data on which analysis needs to be performed is known, it's time to transform it from raw data (.txt) to useable data (.csv) as shown in Fig 4.

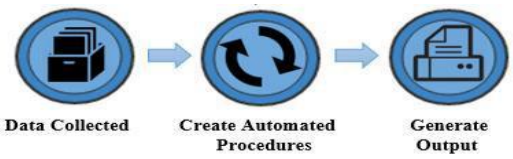

Fig 4: Process involved in Data Transformation

Data cleansing is also known as Data Scrubbing in which inaccurate records from a particular dataset are corrected and eliminated. The purpose of data cleansing is to detect incorrect, irrelevant or insufficient parts of the data to either alter or delete it to ensure that a given set of data is accurate and consistent with other sets in the system.

Validations to be performed on the text file to avoid any issues while reading your file into Python:

- Avoid blank space in names, values or fields, else each word will be treated as a separate variable, resulting in errors that are related to the number of elements per line in your data set.
- To concatenate words, do this by making using of a dot (.) .For example Sender. Age
- Short names are preferred over long names.
- To avoid special symbols such as !, @, #, $, ^, ***, ,(,) ,- , ?,,,< ,> , / , | , \ , [,] ,{ , and }.
- Values missing in the data set are tend to be indicated with NA.

Performing all such validations on text file and hence converting it into csv file via the help of excel proved to be a tedious job and hence automating the process of conversion of text file into csv file by writing few lines of code proved to be more efficient, less time

consuming as well as reduced manual work. The process of conversion with all the above mentioned validations thus shifted from hours to minutes.

5.3 Data Loading

Data loading stage includes importing resultant file in to algorithm as on Fig. 7.

```
Please input chat filepath:1.txt
Please select common word file or leave it blank to escape:
    1: Indonesian (id_cw.py)
    2: English (en_cw.py)
    3: Custom file
    4: Skip common word
2
You wanna print the verbose mode? y/[N]: y
```

Fig 5: WhatsApp Data Load into the algorithms.

```
Extracting data. Please wait....
+++01/07/18, 3:27 pm - +91 78988 00841: Group mai judne ke Baad AAP sabhi apna apna Parichay bhi dijiye
+++01/07/18, 3:50 pm - +91 99267 80970: अशोक कुमार सोनी, कोरबा।
|||SECL में जॉब करता हूँ.
+++01/07/18, 3:52 pm - +91 76919 12195: This message was deleted
+++01/07/18, 3:54 pm - +91 97132 55498: Rahul soni basin se🙏
+++01/07/18, 3:54 pm - +91 76919 12195: मैं आशुतोष सोनी जशपुर में बाल न्यायालय में नौकरी करता हूँ
+++01/07/18, 3:55 pm - +91 98264 52324: जय जोहार
+++01/07/18, 3:56 pm - +91 93290 02307: मैं संदीप सोनी घरघोड़ा जिला रायगढ़ से हु🙏🙏
+++01/07/18, 3:56 pm - +91 78988 00841: Main Vinay Kumar Soni mungeli se bank main Kam karta ho 7898800841
+++01/07/18, 3:58 pm - +91 97132 55498: Aap sb is grup me sadi yogy bayodata bhej skte hai kio ki is grup me sbhi cg se hai
+++01/07/18, 4:00 pm - +91 76919 12195: जी हां
@@@01/07/18, 4:00 pm - +91 98264 52324: <Media omitted>
+++01/07/18, 4:01 pm - +91 88272 72200: Mai Amitesh Soni
|||From Raipur
|||Profession - CA, CS, M. Com, LLM
@@@01/07/18, 4:01 pm - +91 98264 52324: <Media omitted>
+++01/07/18, 4:02 pm - +91 99267 80970: नाम- काजल सोनी
```

Fig 6: Extracting Content From Loaded Data

5.4 Exploratory Data Analysis (EDA) and Visualization

EDA is an approach to data analysis for summarizing and visualizing the important characteristics of a data set [9].

Fig 7: Day Wise Data Analysis and Visualization

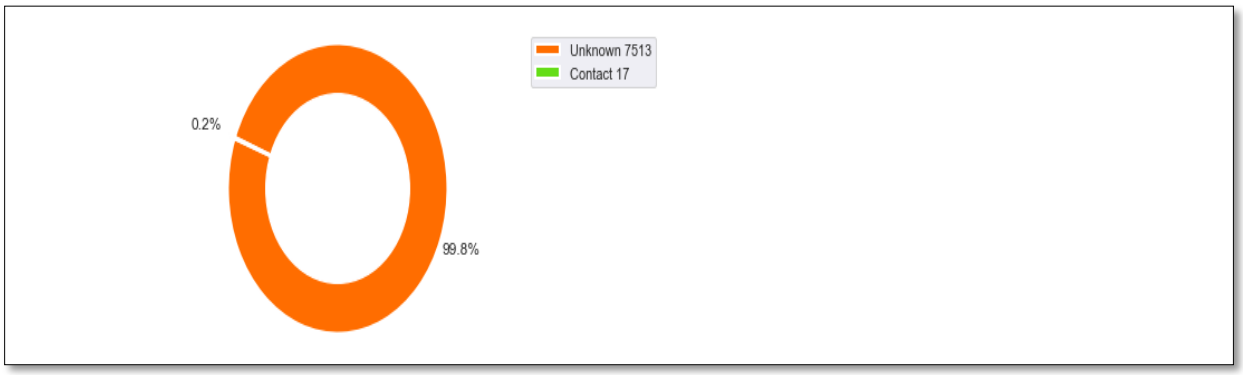

Fig 8: Day Wise Data Analysis and Visualization

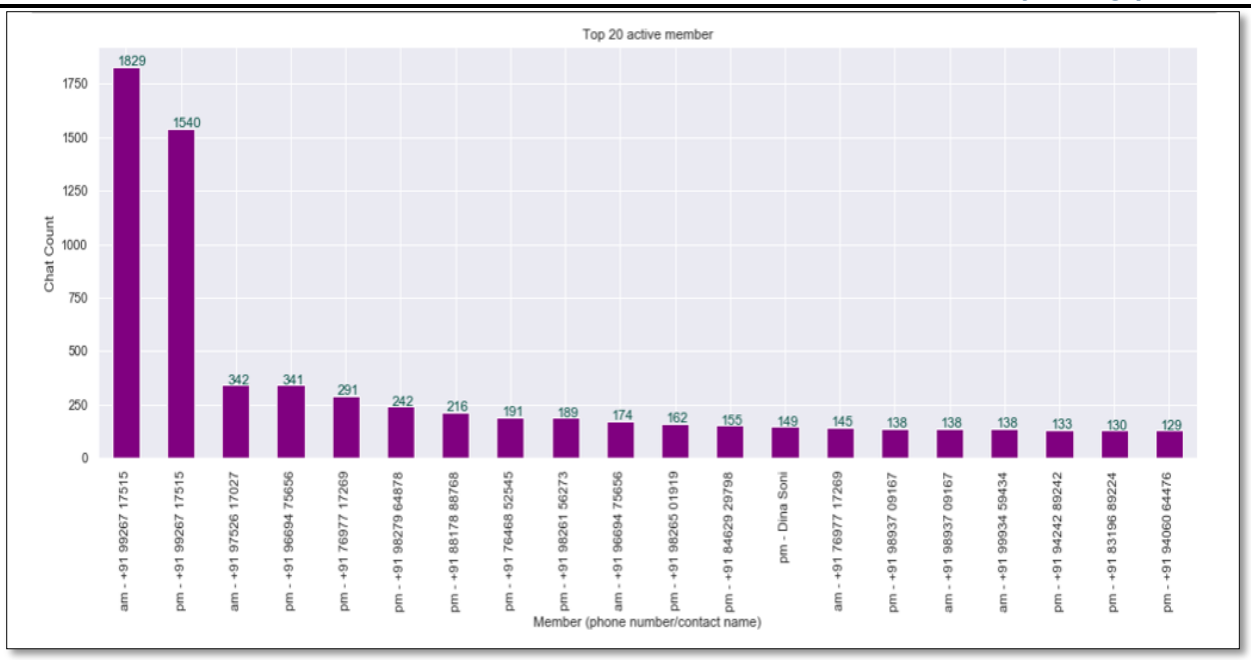

Fig 9: Day Wise Data Analysis and Visualization

Fig 10: Day Wise Data Analysis and Visualization

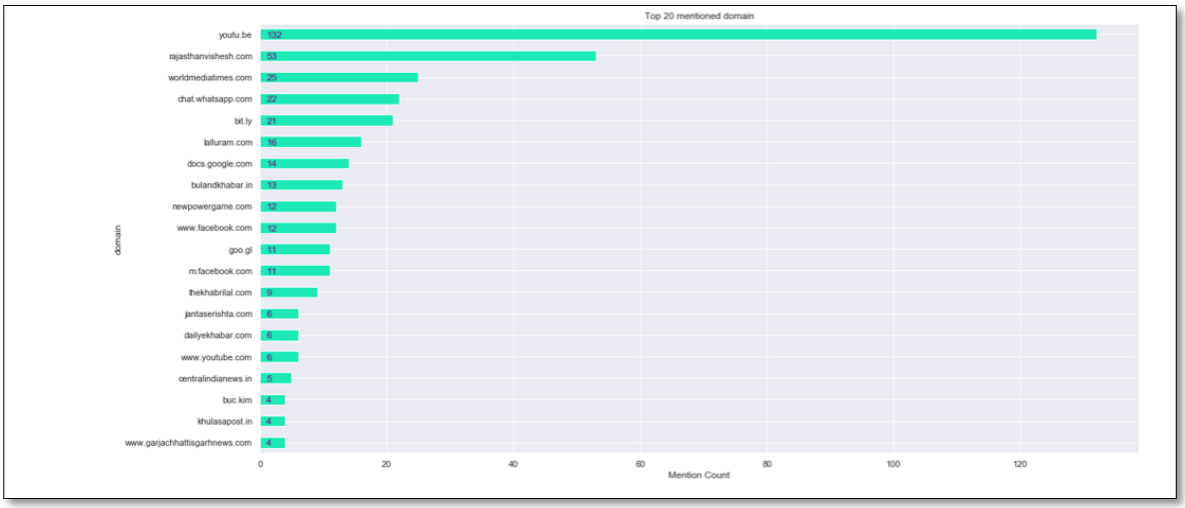

Fig 11: Day Wise Data Analysis and Visualization

```
#Total Chat: 12267
#Deleted Message: 0
#Total Attachment: 7530
  - : ('unknown', 7513)
  - : ('contact', 17)
```

Fig 12: Total Chat Summary

```
#Top 20 Member
-----------------
                          chat_count
member
am - +91 99267 17515        1829
pm - +91 99267 17515        1540
am - +91 97526 17027         342
pm - +91 96694 75656         341
pm - +91 76977 17269         291
pm - +91 98279 64878         242
pm - +91 88178 88768         216
pm - +91 76468 52545         191
pm - +91 98261 56273         189
am - +91 96694 75656         174
pm - +91 98265 01919         162
pm - +91 84629 29798         155
pm - Dina Soni               149
am - +91 76977 17269         145
pm - +91 98937 09167         138
```

Fig 13: Top 20 Reader Result.

```
#Top 20 Words
-----------------
('soni', 533)
('raipur', 432)
('hospital', 397)
('l', 251)
('hai', 196)
('g', 175)
('m', 174)
('c', 166)
('message', 156)
('pm', 133)
('deleted', 126)
('shri', 125)
('dob', 110)
('se', 109)
('n', 107)
('ka', 104)
('bilaspur', 98)
('ke', 97)
```

Fig 14: Top 20 Words Result.

```
#Top 20 Emoji
-----------------
          char_count
emj_char
🙏               3382
〰               1510
🌹                995
🍀                984
👌                533
👍                421
🔸                418
👉                406
😂                341
🌸                339
👇                331
🍀                304
👍                295
🚩                279
💪                241
```

Fig 15: Top 20 imoji Result.

```
#Top 20 Mentioned Website
-----------------
                              d_count
domain
youtu.be                          132
rajasthanvishesh.com               53
worldmediatimes.com                25
chat.whatsapp.com                  22
bit.ly                             21
lalluram.com                       16
docs.google.com                    14
bulandkhabar.in                    13
newpowergame.com                   12
www.facebook.com                   12
goo.gl                             11
m.facebook.com                     11
thekhabrilal.com                    9
jantaserishta.com                   6
dailyekhabar.com                    6
```

Fig 16: Top 20 Mentioned website Result.

```
In KNN Classifier
fscore: 99.79
precision: 99.29
recall: 99.89
Execution Time: 0.01
In Naïve Byes Classifier
fscore: 99.31
precision: 99.86
recall: 99.58
Execution Time: 0.28
In Decision Tree Classifier
fscore: 99.78
precision: 99.37
recall: 99.19
Execution Time: 0.19
In Random Forest Classifier
fscore: 99.55
precision: 99.15
recall: 99.63
Execution Time: 0.29
In SVM Classifier
fscore: 99.88
precision: 99.74
recall: 99.77
Execution Time: 0.12
```

Fig 16: Performance of various classifiers

The aim of this research is to classify the number of users as those addicted and not addicted to WhatsApp group chat and thus predicting the level of addiction as well as to find a way to answer the below questions :

A total of 3674 messages had been de
livered after noon while a whole of 1889 messages were delivered before noon. As a result maximum amount of interactions took place after noon as shown in Fig 17 with the help of Pie chart.

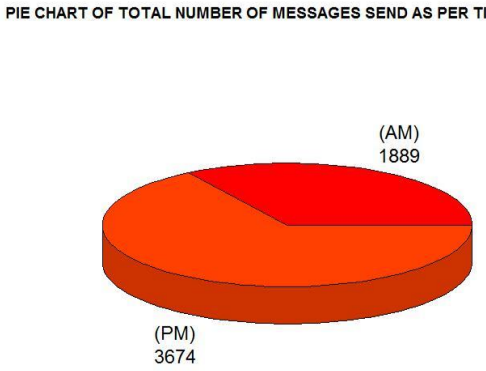

Fig 17: Total Number of Messages send as per timestamp

5.5 Results and Discussion

The classification algorithms K-NN, Naïve Bayes, Decision Tree, Random Forest and SVM have their own importance and we use them on the behavior of the Whatsapp chat log, but on the basis of Table 1 we found that K-NN classification algorithm is simplest algorithm as compared to other algorithms.

```
In KNN Classifier
fscore: 99.79
precision: 99.29
recall: 99.89
Execution Time: 0.01
In Naïve Byes Classifier
fscore: 99.31
precision: 99.86
recall: 99.58
Execution Time: 0.28
In Decision Tree Classifier
fscore: 99.78
precision: 99.37
recall: 99.19
Execution Time: 0.19
In Random Forest Classifier
fscore: 99.55
precision: 99.15
recall: 99.63
Execution Time: 0.29
In SVM Classifier
fscore: 99.88
precision: 99.74
recall: 99.77
Execution Time: 0.12
```

Fig 18: compute performance on varies classifier.

Table 1: Various Classification Results of Whatsapp Chat Log

Algorithm	Accuracy	Pricision	Recall	Execution Time
K-nn	99.9	98.9	99.2	15
Naïve Byes	97.2	98	98.7	14
Decision Tree	98	97	96	20
Random Forest	75	84.64	76	141.6
SVM	98	98.33	98	91.8

VI. CONCLUSION

From the performed analysis and visualization it is found that total number of active users in WhatsApp group chat are 22 consisting of equal number of males and females. Majority of the female users tend to be more addicted to WhatsApp group chat as compared to male users, due to various features provided by WhatsApp such as multimedia, Smiley and Text. The most addicted respondents were in the age group of 20 to 30 years representing a young sample. So as to conclude WhatsApp is one of the best communication platform whose pros and cons are decided by the user itself .If used positively then it's a boon for the users and if addicted then a ban and thus this research paper classified the level of addiction of users to the WhatsApp group chat so as to limit the time spend on it and to explore the group whenever necessary. The implementation results show the values for Accuracy, Precision, Recall and execution time. The overall results for the all the algorithms are shown in Table1. From the results it is clear that, the K-NN algorithm is better than other algorithms on both the datasets.

VII. REFERENCES

[1] D.Radha, R. Jayaparvathy, D. Yamini, "Analysis on Social Media Addiction using Data Mining Technique", International Journal of Computer Applications (0975 – 8887) Volume 139 – No.7, pp. 23-26, April 2016.

[2] Sanchita Patil," Big data analytics using R**",** International Research Journal of Engineering and Technology (IRJET) Volume 3, Issue 7 July 2016,pp. 78-81.

[3] Sagar Deshmukh, "Analysis of WhatsApp Users and Its Usage worldwide", International Journal of Scientific and Research Publications, Volume 5, Issue 8, pp. 1-3,August 2015 1 ISSN 2250-3153.

[4] https://en.wikipedia.org/wiki/Data_analysis#cite_note-O.27Neil_and_Schutt_2014-3

[5] https://www.reference.com/education/purpose-collecting-data-8d8be32cc477eb45#

[6] Tal Galili, "R-bloggers", December 10, 2015. [Online] Available: http://www.r-bloggers.com/how-to-learn-r-2/ [Accessed: 23-July- 2016]

[7] SSCC (social science computing cooperative), "R for Researchers:Projects". [Online]Available:http://www.ssc.wisc.edu/sscc/pubs/RF R/RFR_Projects.html [Accessed: 23- July- 2016]

[8] "This R Data Import Tutorial Is Everything You Need",July 21st, 2015 in R Programming. [Online]Available:https://www.datacamp.com/communit y/tutorials/r-data-import-tutorial#gs.EYeqhvc .[Accessed: 23- July-2016]

[9] Jovial, "Exploratory data analysis with R", [Online] Available: https://rpubs.com/Jovial/R [Accessed: 23-July- 2016]

[10] Abhishek Kaushik and Sudanshu Naithani "A Comprehensive study of Text Mining Approach".

[11] Vishal Gupta And Gurpreet S. Lehal "A Survey of Text Mining Techniques and Application".

[12] Ms. Anjali Ganesh Jivani, Prof. B. S. Parekh "A Comparative Study of Text Data Mining Algorithms and its Applications".

[13] S.Niharika,V.SnehaLatha and D.R.Lavanya"A Survey On Text Categorization".

[14] D. E. Johnson, F. J. Oles, T. Zhang, T. Goetz,"A decision-tree-based symbolic rule induction system for text categorization", IBM Systems Journal, September 2002.

[15] Kim S. B., Rim H. C., Yook D. S. and Lim H. S., "Effective Methods for Improving Naïve Bayes Text Classifiers", LNAI 2417, 2002, pp.414-423.

[16] Klopotek M. and Woch M., "Very Large Bayesian Networks in Text Classification", ICCS 2003, LNCS 2657, 2003, pp. 397-406.

[17] Joachims, T., Transductive inference for text classification using support vector machines. Proceedings of ICML-99, 16th International Conference on Machine Learning, eds. I. Bratko & S. Dzeroski, Morgan Kaufmann Publishers, San Francisco, US: Bled, SL, ,1999 ,pp. 200–209.