Mining Social Media Data Using R and WEKA Tools

B. Jayaram

Abstract--- *These days there are lot of data are available in social media in various types. These data can be extracted from various sources using different application software. In this paper I have organized my discussion as follows first part deals about introduction to social media processing, second part explains how social media and machine learning are related with a detailed discussion, third section deals about tools of machine learning tools (WEKA and R) and their comparison, fourth section deals about how to extract data from social mining websites, and the last part explains about how to process social websites data with data big data tools (WEKA and R).*

Keywords--- *Big Data, Social Media, Machine Learning, Data Extraction.*

I. INTRODUCTION

Social media plays a vital role in every human life due to the introduction of lot of gadgets (mobile phone, laptop etc) with lot of facilities in inside those gadgets makes people more interested in social media usage. Social media has become an integral part of human life through various websites and application using mobile applications etc. Most traditional online media include social components, such as comment fields for users. In business, social media is used to market products, promote brands, and connect to current customers and foster new business [1].

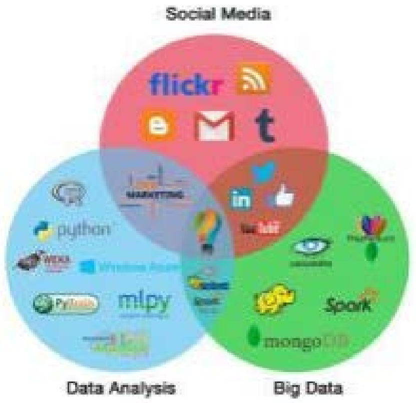


Fig. 1: The framework of bid data and social media. [1]

In general any social media data will consist of three parts big data, social media websites and data analysis part. Figure 1 shows the combination and the interaction between the three modules. Social media is also a part of big data where collection of data and processing requires the following points to be addressed efficiently [4].

The basic 4V which we normally relate to big data is listed below:

B. Jayaram, Assistant Professor, CSE Department, Malla Reddy Institute of Technology, Hyderabad.

Received: 14 Nov 2018 | Revised: 23 Dec 2018 | Accepted: 06 Jan 2019

- Volume:
    - It deals about the storage space required for the data processing.
- Velocity:
    - It deals with how fast the data can be collected and how quick the efficient output is produced.
- Variety:
    - The data can be in various forms, it may be structured or unstructured specified in a certain format of storage.
- Veracity:
    - It deals with uncertainty in the data quality.

The main steps for social media analytics are listed below [4].

- Discovery: Discovering the structure and patterns.
- Tracking: involved decision on selecting a data source from various sources Eg: Twitter, Facebook, etc.
- Preparation: This involves preprocessing of data before processing.
- Analysis: This is used for analyzing the data based on purpose of data.

The data in social media can be classified into following categories with examples given as below.

- **Blogs:** Blogger, LiveJournal, WordPress
- **Microblogs:** Twitter, Google Buzz
- **Opinion mining:** Epinions, Yelp
- **Photo and video Sharing:** Flickr, YouTube
- **Social bookmarking:** Delicious, StumbleUpon
- **Social networking sites:** Facebook, LinkedIn, Myspace, Orkut
- **Social news:** Digg, Slashdot
- **Wikis:** Scholarpedia, Wikihow, Wikipedia, Event maps

II. HOW MACHINE LEARNING WORKS

Machine learning is the scientific study of algorithms and statistical models that computer systems use in order to perform a specific task effectively without using explicit instructions, relying on patterns and inference instead. Machine learning algorithms working can be classified into the following steps.

- Gathering data from various sources.
- Cleaning the data.
- Model Building.
- Gaining knowledge from models
- Visualization of data

Here each step is broadly classified in to a number of sub steps and detailed explanation is mentioned below.

A. Gathering of data from various sources

Data collection can be classified in to three broad categories namely 1) data acquisition 2) data labeling and 3) existing data [2].

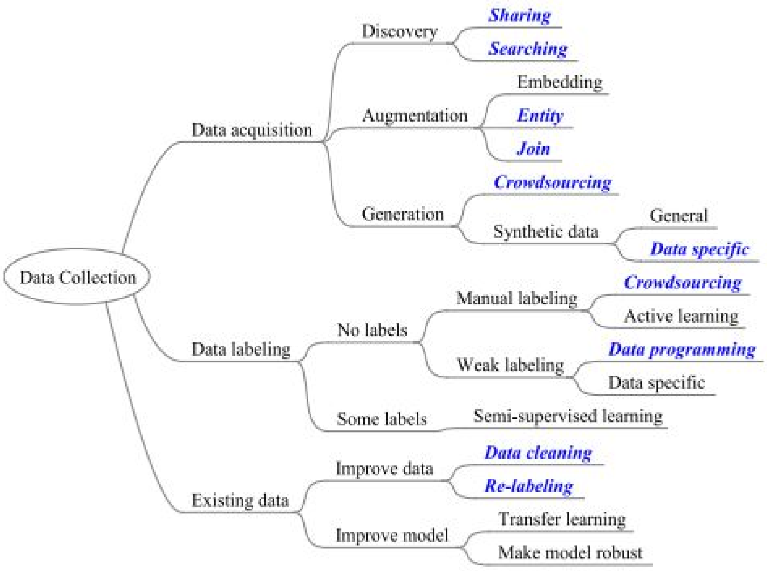


Fig. 2: Data collection for machine learning [2]. Here some of topics mentioned (blue color italics) are least concerned

Data acquisition is a process to find datasets that can be used to train learning models and there are three different approaches for this process [2]. They are

- Data discovery.
- Data augmentation.
- Data generalization.

Data discovery is a process of discovering the data from any source. This can be done 2 different approaches such as,

- Sharing of data using collaborative analysis using data hubs.
- Sharing of data using collaborative web based analysis. Eg: google forums, spreadsheets etc.

Data augmentation adds value to base data by adding information derived from internal and external sources within an enterprise. Data augmentation can be applied to any form of data, but may be especially useful for customer data, sales patterns, product sales, where additional information can help provide more in-depth insight [5].

Some common techniques in data augmentation include the following process [5].

- Extrapolation techniques using heuristics.
- Tagging techniques using grouping.
- Aggregation technique using mathematical calculations.

Received: 14 Nov 2018 | Revised: 23 Dec 2018 | Accepted: 06 Jan 2019

- Probability techniques using probabilities of event occurrences.

Data generation can be done from synthetic data from different points and sources by using any of the following 3 techniques [3].

- Machine - sensors and instruments
- Human - social media and e-mails
- Organization - ERP and other enterprise applications

Data Labeling is an interlinked process with data acquisition since it starts only after data gathering is completed. Data labeling can be done by two different methods

- Using existing labels.
- Using crowd based labels.

Both these methods follow the process of machine learning algorithm classification and regression using semi supervised learning [2].

Existing data can be used to train models in machine learning by two ways. First by improving the data using data cleaning techniques described in next section and data labeling techniques, Second by improving the model by using various learning models [2].

B. Cleaning data to have homogeneity

It is a process of dealing with missing values and deciding what can be done with outliers [5]. So to produce effective prediction missing data can be replaced by the following set of procedures [5].

- Deleting rows that have missing data.
- Replacing missing data with mean or median or mode.
- Assigning a unique category.
- Predicting the missing values.
- Using algorithms which support missing values.

Outliers are a part of any multidimensional data which lies a distance away from any consideration for the model and the reason for this category will probably be poor collection of data [1].

C. Model Building - Selecting the Right machine learning algorithm

This step can be broadly classified in to the following steps [5]:

- Know your data.
- Categorize the problem.
    - Categorize by input.
    - Categorize by output.
- Find the available algorithms.

For **knowing our data** we need to look at the summary statistics and visualizations (Eg: Percentages, average, median etc.). In case of **categorization by input** it means the following is applicable.

- Supervised learning algorithm is followed if we have a labeled data.\
- Unsupervised learning algorithm is followed if we have an un-labelled data and we want to find its structure.
- Reinforcement learning problem is followed if the objective function is to be minimized.

In case of **categorization by output** the following is applicable.

- If the output is a model it is identified as regression problem.
- If the output of the model is set into input groups then it is referred as clustering problem.
- If the output of the model is a class then it will be a classification problem.
- If any anomaly has to be detected then it is a problem of anomaly deduction.

Constraints to be considered are:

- Data storage capacity.
- Speed of prediction.
- Does the learning have to be fast?

D. Gaining knowledge from model result

For gaining **knowledge** from the model then we can use the help any of machine learning tools.

These tools help the user to gain knowledge from the output. Since they have following main advantages [5].

- No programming required to work with the tool. Since they are automatic.
- Better management of work.
- Produces result faster than human processing.
- Produces better quality of result.
- It has better uniformity.

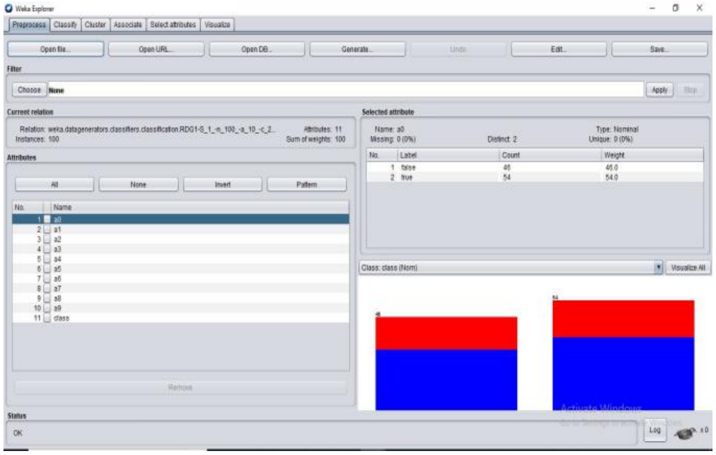
E. Visualization – Transforming results in visual representation

For visualizing the result in an understandable form there are lot free open source tools available in market for machine learning tools [5]. A few of them include

- Weka.
- R Programming.
- Python.
- Apache spark.
- Sas.
- Rapid miner.
- KNIME.
- Splunk.

- QLinkView.
- Orange.

Sample visualization is given using weka tool is shown below in figure 3. This tool helps the user to select the corresponding data set and generate whatever is the need for the end user.



Filter: None

Relation: weka.datagenerators.classifiers.classification.RDG1-S_1_-n_100_-a_10_-c_2... Attributes: 11
Instances: 100 Sum of weights: 100

Name: a0 Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	false	46	46.0
2	true	54	54.0

No.	Name
1	a0
2	a1
3	a2
4	a3
5	a4
6	a5
7	a6
8	a7
9	a8
10	a9
11	class

Class: class (Nom)

Status: OK

Fig. 3: A Sample Screen from WEKA Tool

Thus helps the end user to understand the system easily with the help of representation in graphs. This tool mainly follows supervised and unsupervised learning models.

III. Machine Learning Tools

As discussed in previous section there are a lot of tools available for processing big data and majority of them are open source and free to use under GNU License. In this section a brief discussion about WEKA and R-Programming are given.

A. Weka

Waikato Environment for Knowledge Analysis (**Weka**) is a suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License. Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with command line interface and also graphical user interfaces for easy access to these functions.

B. R Programming

R is a programming language and free software environment for statistical computing supported by R foundation for statistical computing. It is widely used for statistical computing, datamining surveys and data analysis. As on June 2019 R ranks 22nd in the TIOBE index a measure of popularity of programming languages. A GNU package, source code of R software environment is written in C, FORTRAN and R itself and it is freely available under GNU General Public License. Similar to Weka R has both command line and graphical user interface such as RStudio, and it is referred as an IDE (Integrated Development Environment).

Received: 14 Nov 2018 | Revised: 23 Dec 2018 | Accepted: 06 Jan 2019

C. Comparison of WEKA and R

Comparison of WEKA and R-Programming is also discussed here in detailed tables given below. The comparison is can be done based on various factors [6]

- Platform.
- Input – output formats.
- Visualization techniques.

The platform supported for processing big data is shown in table 1 for the above mentioned tools [6].

Table 1: Comparison on WEKA and R based on platform [6]

Platform	**R**	**WEKA**
Windows	Yes	Yes
Mac	Support Mac Classic ended with R 1.7.1	Yes
Unix/Linux	Yes	Yes
Minimum JRE Required	No relation with JRE	1.3
X86 – X64	Yes	Yes
Multi-cores	Yes	Yes
Distributed computing	Yes	Yes
Client - Server	Yes	Yes

Comparison based file format supported for input and output is listed below in table 2.

Table 2: File format supported in WEKA and R [6]

Format	**WEKA**	**R**
CSV	Yes	Yes
ARFF	Yes	Yes
C4.5 Format	Yes	No
database	Yes	Yes
SAS	No	Yes
SPSS	No	Yes
Minitab	No	Yes

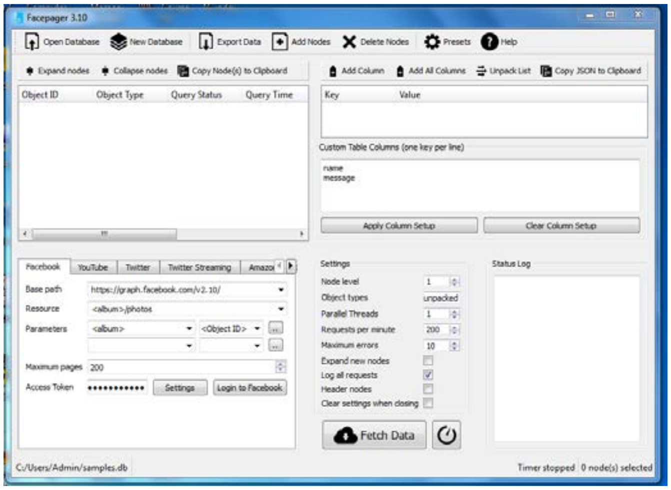
Comparison based on visualization techniques is listed below in table 3[6]. Here data can be represented for analysis using various aspects of processing. Visualization is generally a process of representing data in understandable manner for the user.

Table 3: Comparison on WEKA and R for visualization [6]

Feature	WEKA	R
Descriptive statistics	Yes	Yes
Frequency table	Yes	Yes
Scatter plot	Yes	Yes
Scatter plot matrices	Yes	Yes
Histogram	Yes	Yes
Tree / graph visualization	Yes	Yes
Boxplot	No	Yes
ROC Curve	Yes	Yes
Precision / Recall curve	Yes	Yes
Life chart	Yes	Yes
Cost curve	Yes	Yes

IV. Extraction Of Data From Social Data Mining Websites

There are a lot of free tools available online for extracting data from social data mining websites. In this paper I have used **Face Pager version 3.10** [7]. This software supports data extraction from various social mining websites such as Facebook, Twitter, Amazon, YouTube etc. Face pager was made for fetching public available data from Facebook, Twitter and other JSON-based APIs. All data is stored in a SQLite database and may be exported to csv. It supports various platforms windows, MAC and Linux environments. To fetch data from the social mining websites first we have to pass the object id as input, object id can be fetched by logging into our social networking login and passing the resource and parameter to fetch data.



Custom Table Columns (one key per line): name
message

Base path: https://graph.facebook.com/v2.10/
Resource: <album>/photos
Parameters: <album> <Object ID>
Maximum pages: 200
Access Token: ●●●●●●●●●●●

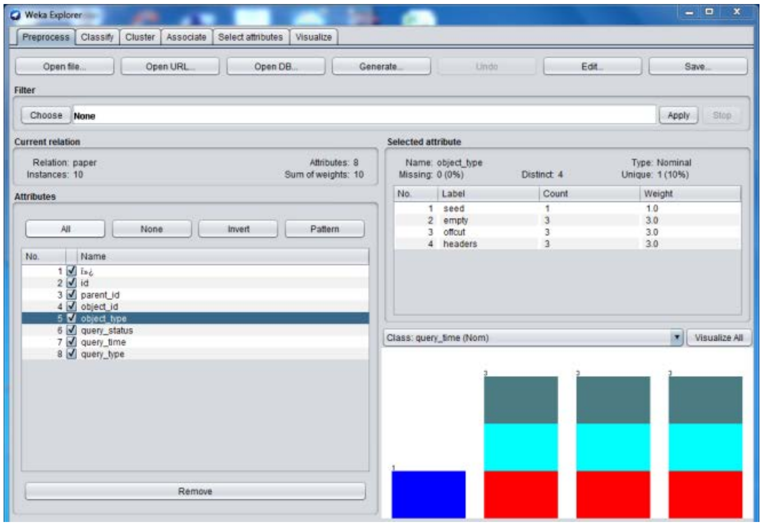
Settings
Node level: 1
Object types: unpacked
Parallel Threads: 1
Requests per minute: 200
Maximum errors: 10
Expand new nodes ☐
Log all requests ☑
Header nodes ☐
Clear settings when closing ☐

C:/Users/Admin/samples.db
Timer stopped | 0 node(s) selected

Fig. 4: Sample Screenshot of Face Pager Version 3.10

Received: 14 Nov 2018 | Revised: 23 Dec 2018 | Accepted: 06 Jan 2019

Eg: from Facebook if I want to extract friend's data in my login. I have to pass website as Facebook and login to my Facebook login in Face Pager software and pass object to extract data as friends. After successful attempt corresponding data can be fetched from the corresponding social mining website and stored in various formats such as CSV and JSON. Figure 4 shows the sample screenshot of Face pager version 3.10**.** For comparison purposes in this paper I have extracted details from my Facebook login and selected the resource as **The Hindu – Tamil** and extracted a sample data with 8 attributes and 4 rows of data and generated a CSV file as extracted data. When this CSV file is given as input to WEKA tool it shows the following sample screenshot for a particular attribute as shown in figure 5.



Filter: None

Current relation

Relation: paper
Instances: 10
Attributes: 8
Sum of weights: 10

Selected attribute

Name: object_type
Missing: 0 (0%)
Distinct: 4
Type: Nominal
Unique: 1 (10%)

No.	Label	Count	Weight
1	seed	1	1.0
2	empty	3	3.0
3	offcut	3	3.0
4	headers	3	3.0

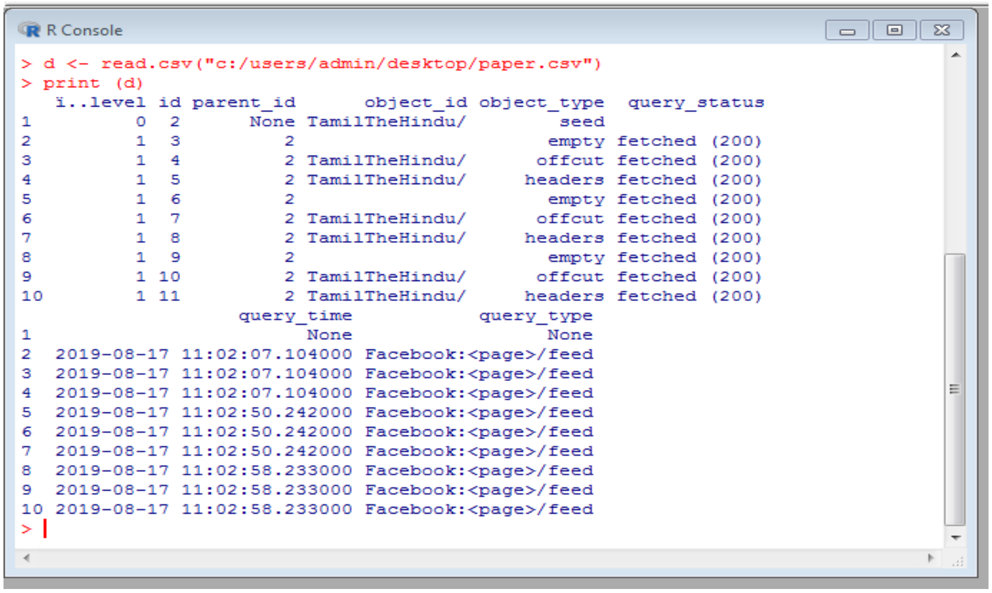
Attributes

No.		Name
1	☑	ï»¿
2	☑	id
3	☑	parent_id
4	☑	object_id
5	☑	object_type
6	☑	query_status
7	☑	query_time
8	☑	query_type

Class: query_time (Nom)

Fig. 5: WEKA screenshot for Object Type attribute on The Hindu – Tamil

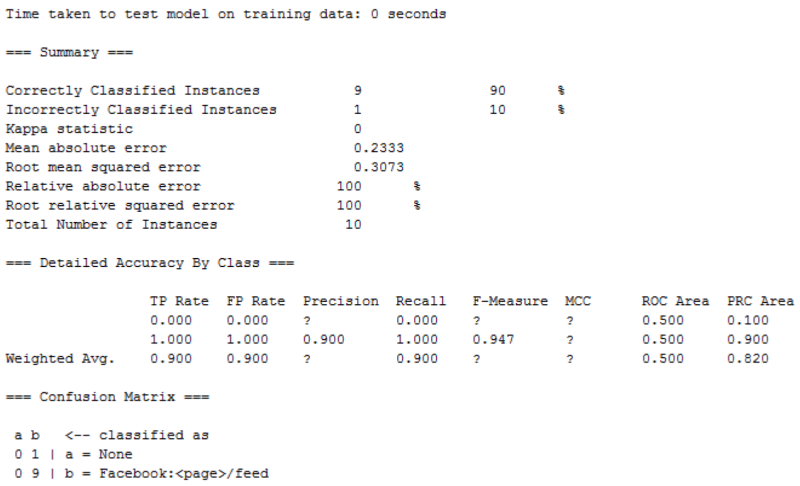
Similarly using R-studio I can read the same CSV file and get following output as shown below in figure 6.



```
> d <- read.csv("c:/users/admin/desktop/paper.csv")
> print (d)
   i..level id parent_id      object_id object_type  query_status
1         0  2      None TamilTheHindu/        seed
2         1  3         2                      empty fetched (200)
3         1  4         2 TamilTheHindu/      offcut fetched (200)
4         1  5         2 TamilTheHindu/     headers fetched (200)
5         1  6         2                      empty fetched (200)
6         1  7         2 TamilTheHindu/      offcut fetched (200)
7         1  8         2 TamilTheHindu/     headers fetched (200)
8         1  9         2                      empty fetched (200)
9         1 10         2 TamilTheHindu/      offcut fetched (200)
10        1 11         2 TamilTheHindu/     headers fetched (200)
                   query_time           query_type
1                        None                 None
2  2019-08-17 11:02:07.104000 Facebook:<page>/feed
3  2019-08-17 11:02:07.104000 Facebook:<page>/feed
4  2019-08-17 11:02:07.104000 Facebook:<page>/feed
5  2019-08-17 11:02:50.242000 Facebook:<page>/feed
6  2019-08-17 11:02:50.242000 Facebook:<page>/feed
7  2019-08-17 11:02:50.242000 Facebook:<page>/feed
8  2019-08-17 11:02:58.233000 Facebook:<page>/feed
9  2019-08-17 11:02:58.233000 Facebook:<page>/feed
10 2019-08-17 11:02:58.233000 Facebook:<page>/feed
> 
```

Fig. 6: Data Represented R – Studio

Received: 14 Nov 2018 | Revised: 23 Dec 2018 | Accepted: 06 Jan 2019



```
Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances           9               90      %
Incorrectly Classified Instances         1               10      %
Kappa statistic                          0
Mean absolute error                      0.2333
Root mean squared error                  0.3073
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances               10

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.100
                 1.000    1.000    0.900      1.000    0.947      ?        0.500     0.900
Weighted Avg.    0.900    0.900    ?          0.900    ?          ?        0.500     0.820

=== Confusion Matrix ===

 a b   <-- classified as
 0 1 | a = None
 0 9 | b = Facebook:<page>/feed
```

Fig. 7: Confusion matrix of the data

Classification of data using weka gives the following confusion matrix as below in figure 7. For plotting of data in R we get the figure 8 as resultant plot for all attributes.

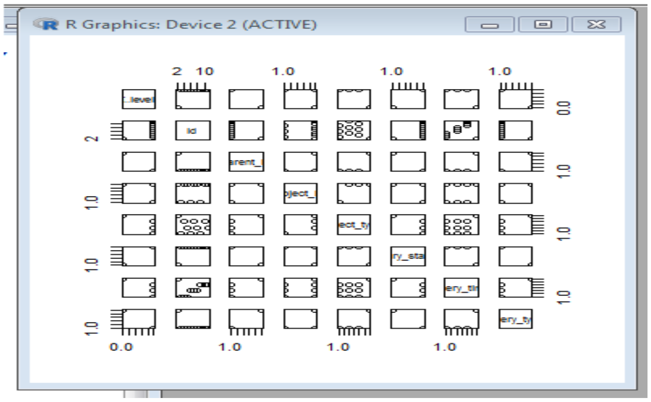


Fig. 8: Plotting using R – GUI for all attributes

Similarly plotting using weka for all attributes gives the figure 9 as output. Here we see that all nominal type attributes are plotted.

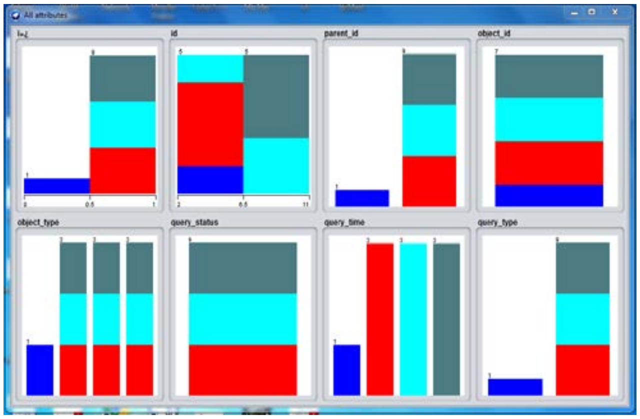


Fig. 9: Plotting using WEKA for all attributes

V. CONCLUSION

In this paper we can understand the relation between big data and social data mining, and how to extract data from the mining websites using tools available online. The performance can be measured for social data mining websites through the use of existing tools by using the data extracted and perform any analysis required.

The future scope of this paper is to work with a larger data set that is extracted from social mining websites and compare various features for analysis purpose and also work with various types of data to produce effective output at an efficient time.

REFERENCES

[1] Jayaram, B., Jeyachandran, A., Dileep, M., & Joseph Shastry, K.S.S. (2018). A Survey On Social Media Data Analytics And Cloud Computing Tools. *International Journal of Mechanical and Production Engineering Research and Development*, *8*(3), 243-254.

[2] Roh, Y., Heo, G., & Whang, S. E. (2019). A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, 1-20.

[3] Ghotkar, M., & Rokde, P. (2016). Big Data: How it is Generated and its Importance. *In Proceedings of National Conference on Recent Trends in Computer Science and Information*, 1-5.

[4] Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. *International journal of information management*, *39*, 156-168.

[5] Jayaram, B., & Kalimuthu, M. (2019). A Study on Machine Learning and Its Workings. *International Journal of Modern Trends in Science and Technology*, *5*(7), 44 – 48.

[6] Al-Khoder, A., & Harmouch, H. (2015). Evaluating four of the most popular open source and free data mining tools. *Int. J. Acad. Scient. Res*, *3*(1), 13-23.

[7] Salloum, S.A., Al-Emran, M., & Shaalan, K. (2017). Mining social media text: extracting knowledge from Facebook. *International Journal of Computing and Digital Systems*, *6*(02), 73-81.

Received: 14 Nov 2018 | Revised: 23 Dec 2018 | Accepted: 06 Jan 2019