Big Data and Web : DISC in Web Analytics For Large Data Popularity of Online News on Mapreduce Clusters

Mahesh S Nayak
Research and Development Centre
Bharathiar University,
Coimbatore – 641 046

Dr. M. Hanumanthappa
Professor, Department of Computer
Science & Applications,
Bangalore University, Bangalore.

Dr. S.Kavitha,
Asst.Professor
Dayananda Sagar College,
Bengaluru – 560078

Abstract— Big data is gathering and analyzing the data. Without analytics, it's just a bunch of data with limited business use. The big amounts of data storing from various sources the significance of analytics has enormously grown making the companies to tap the bunch of data that was considered useless all these years. The importance of big data is given preference as bound to provide results on the fly. The MapReduce paradigm has long been a staple of big data computational strategies. With the development of the Internet with website, people communicate online news articles every day. The percentage of message communication any news article indicates how popular the news is. The objective of the paper is to find the best model and set of feature to predict the popularity of online news, using machine learning techniques. The dataset is collected from UCI Mashable online news website. For the feature set preprocessing is done using unsupervised learning method AddExpression-E 0.0 expression algorithm. We have implemented two classifiers using ZeroR and RepTree classification. We have successfully implemented two clustering algorithms KMeans clustering and Canopy cluster algorithm. Their performances are recorded and compared. The ZeroR gives the best result with less time complexity of 0.05 seconds and canopy clustering took 2.09 seconds. The research can be used in any organization to choose the classifier and cluster algorithm for process the dataset.

Keywords – Big Data, Web, Machine learning; Classification; Clusters

1.Introduction

A.Big data

The sets of data that are large in volume is known as Big Data. The data sets can be mined using big data which includes unstructured, semistructured, and unstructured data. It is very difficult to capturing, managing, and processing the data as complex with less time complexity. The big data plays an important role because of an extreme volume of data and a broad variety of types of data.

Data analytics tools are used to analyze the huge data sets. Without analytics, it's just a mass of data with only for restricted business use. After applying the big data analytics the uses like improvement in efficiency, good customer services, sales improvement, with best efficiency. Data analytics involves investigate the data sets to increase insights and decide conclusions such as predicted value for the upcoming activity.

Big data analytics is useful in identifying patterns and relationships in data set and also to apply various statistical techniques to check whether an hypothesis of the given data set is true or not.

B. Online News Communication

In the daily life in current era is addicted on social communications like online communication to friends through social networks using mobiles or computers. Online communication is helpful in online shopping, searching the particular topic, online payment of bills, e-education and e-banking. When all are started doing online activities as mentioned, it would be greatly helpful if we could accurately predict the popularity of news prior to its publication.

The objective of the paper is to collect the big data set online news popularity and analyze the different classification and clustering algorithms. In this research sparse literature survey is found viz., Ranking SVMs [6], Naive Bayes [5] are investigated, and more advanced algorithms such as Random Forest, Adaptive Boosting [4] could increase the precision, analyzing early users' comments [3], or features about post contents and domains [5].

2. Data Set Information:

A. The online news dataset is collected by UCI machine learing website. It summarizes a different group of features regarding articles published by Mashable for two years. The data set is used to classification algorithms and clustering algorithms to predict the best algorithm to find popularity of news articles in big data set.

B. Attribute Information:

The dataset is multivariate characteristics[2]. The dataset contains integer and real, totally 39797 instances, 61 attributes.The full feature set is visualized in the figure 1.

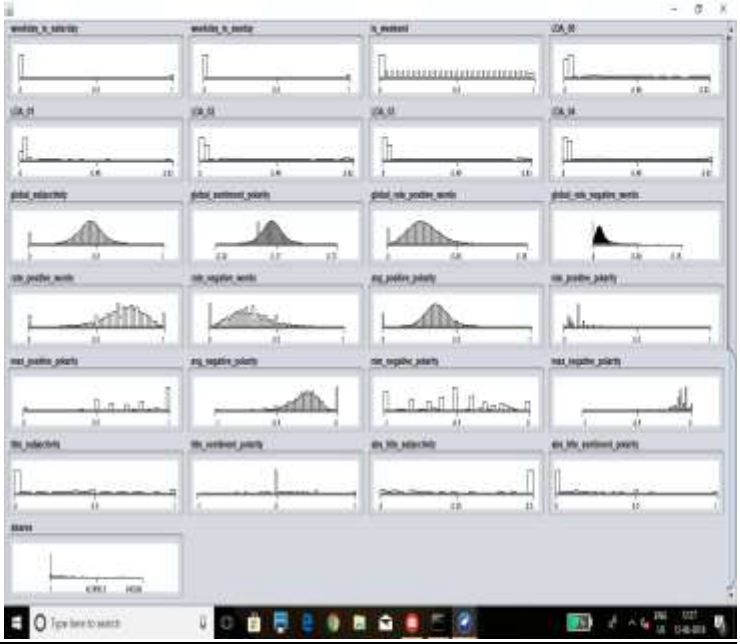
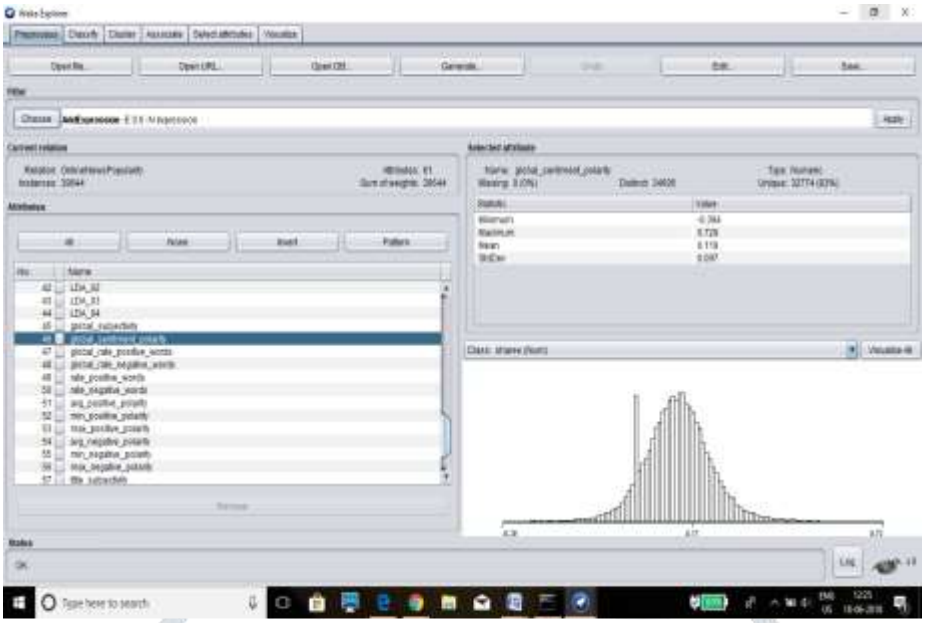


Fig. 1. 61 Attributes visualization

C . Feature set

For the entire feature set preprocessing is done using unsupervised learning method AddExpression-E 0.0 expression algorithm.(Fig 2). The AddExpression applys a mathematical expression involving attributes and numeric constants to a dataset. A new attribute is appended after the last attribute that contains the result of applying the expression.The operators which are supported for the algorithm are: +, -, *, /, ^, log, abs, cos, exp, sqrt, floor, ceil, rint, tan, sin, (,). The –E filter option specifies the expression to apply.



Filter: AddExpression -E 0.0 -N expression

Relation: OnlineNewsPopularity
Instances: 39644
Attributes: 61
Sum of weights: 39644

Name: global_sentiment_polarity
Missing: 0 (0%)
Distinct: 34606
Type: Numeric
Unique: 32774 (83%)

Statistic	Value
Minimum	-0.394
Maximum	0.728
Mean	0.119
StdDev	0.097

No.	Name
42	LDA_02
43	LDA_03
44	LDA_04
45	global_subjectivity
46	global_sentiment_polarity
47	global_rate_positive_words
48	global_rate_negative_words
49	rate_positive_words
50	rate_negative_words
51	avg_positive_polarity
52	min_positive_polarity
53	max_positive_polarity
54	avg_negative_polarity
55	min_negative_polarity
56	max_negative_polarity
57	title_subjectivity

Class: shares (Num)

Status: OK

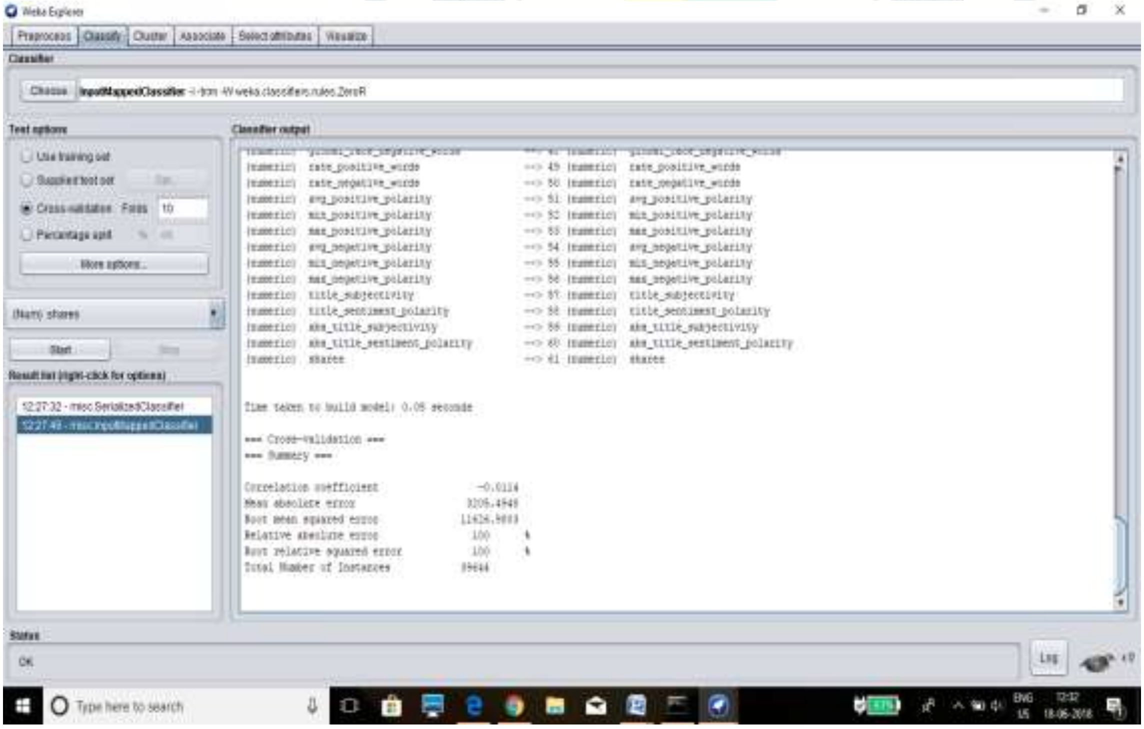
Fig 2: Preprocessing method AddExpression-E 0.0 expression

3. Resulta and Discussion

A.Classification

i.ZeroR

Classification using misc-InputMappedClassifier-I-Trim –w weka.Claasifiers.rules.zeroR algorithm. ZeroR is the simplest classification method which relies on the target and ignores all predictors. ZeroR classifier simply predicts the majority category (class). Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods.The zeroR algorithm constructs a frequency table for the target and select its most frequent value shown in fig 3.



Classifier: InputMappedClassifier -I -trim -W weka.classifiers.rules.ZeroR

Test options: Cross-validation Folds 10

(Num) shares

Result list:
12:27:32 - misc.SerializedClassifier
12:27:48 - misc.InputMappedClassifier

```
(numeric)  global_rate_negative_words        --> 48 (numeric)  global_rate_negative_words
(numeric)  rate_positive_words               --> 49 (numeric)  rate_positive_words
(numeric)  rate_negative_words               --> 50 (numeric)  rate_negative_words
(numeric)  avg_positive_polarity             --> 51 (numeric)  avg_positive_polarity
(numeric)  min_positive_polarity             --> 52 (numeric)  min_positive_polarity
(numeric)  max_positive_polarity             --> 53 (numeric)  max_positive_polarity
(numeric)  avg_negative_polarity             --> 54 (numeric)  avg_negative_polarity
(numeric)  min_negative_polarity             --> 55 (numeric)  min_negative_polarity
(numeric)  max_negative_polarity             --> 56 (numeric)  max_negative_polarity
(numeric)  title_subjectivity                --> 57 (numeric)  title_subjectivity
(numeric)  title_sentiment_polarity          --> 58 (numeric)  title_sentiment_polarity
(numeric)  abs_title_subjectivity            --> 59 (numeric)  abs_title_subjectivity
(numeric)  abs_title_sentiment_polarity      --> 60 (numeric)  abs_title_sentiment_polarity
(numeric)  shares                            --> 61 (numeric)  shares


Time taken to build model: 0.05 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                 -0.0114
Mean absolute error                   3205.4948
Root mean squared error              11626.9883
Relative absolute error                 100      %
Root relative squared error             100      %
Total Number of Instances             39644
```

Status: OK

Fig 3: Cross validation summary of ZeroR

The scheme used is weka.classifiers.misc.InputMappedClassifier -I -trim -W weka.classifiers.rules.ZeroR. Totally 39644 instances, 61 Attributes,10-fold cross-validation Test mode used. Full training set is used for the classification. Using the InputMappedClassifier ZeroR predicts class value of 3395.3801836343455.Time taken to build model: 0.05 seconds.

Correlation coefficient	**-0.0114**
Mean absolute error	**3205.4948**
Root mean squared error	**11626.9803**
Relative absolute error	**100 %**
Root relative squared error	**100 %**
Total Number of Instances	**39644**

ii.REPTree classifier

REPTree is the fast decision tree learner compared to other classifiers shown in fig 4.The resultant of the tree is decision or regression tree using information gain or variance[1]. The algorithm improves the performance with less error with backfitting[7]. It arranges the data points for numeric attributes as shown in the diagram.

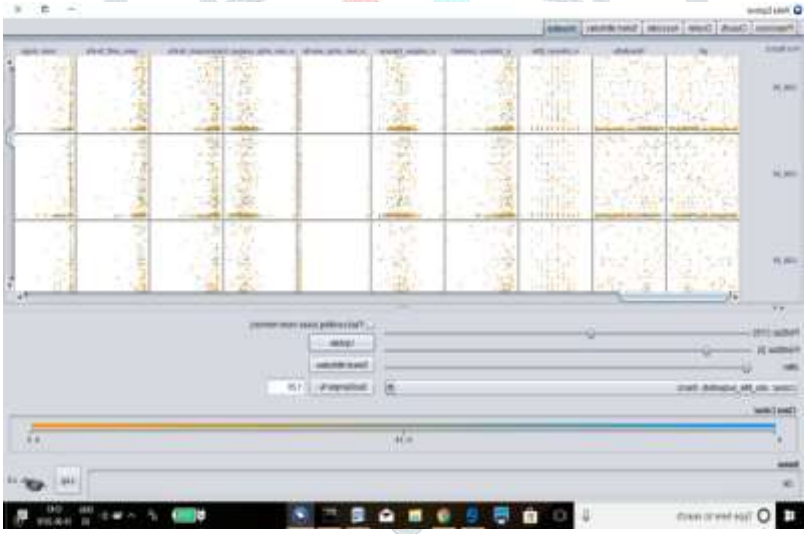


Fig 4: Result of REPTree classifier

Time taken to build model: 0.91 seconds.

B. Clustering

i. kMeans clustering

Total number of 19 iterations.Within cluster sum of squared errors: 154201.3671388387.

Time taken to build model (full training data) : 2.3 seconds

=== Model and evaluation on training set ===

Clustered Instances 0 20933 (53%) 18711 (47%)

ii. Canopy Clustering Algorithm

The Canopy algorithm basically uses mathematical distance functions,hence it needs thresholds as its applicability for multi-dimensional. The canopy clustering algorithm belongs to unsupervised pre-clustering algorithm.The preprocessing stage for the K-means algorithm is done by Canopy algorithm. As the result of less time complexity, this algorithm is applied on many large datasets for evaluation and predictions(fig 5).

Start with the set of data points to be clustered.n=number of points in dataset

Step 1: Start with new canopy after deleting the point from the dataset
Step 2: while(n>0)
For the remaining points, if the distance<L is assign it to the new canopy L=loose distance
Step 3: distance<=tight distance, delete it

Where all the data points are not belongs to the same canopy. The advantages of the K-Means algorithm is that The number of instances of training data that must be compared at each step is reduced and there is some evidence that the resulting clusters are improved. Table 1 shows the total number of cluster with cluster 0 and cluster 1.

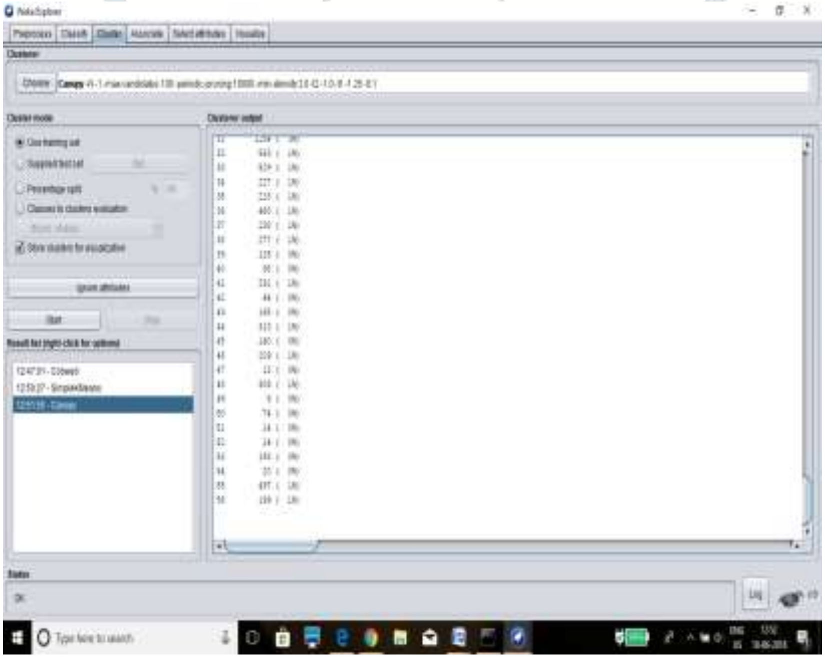
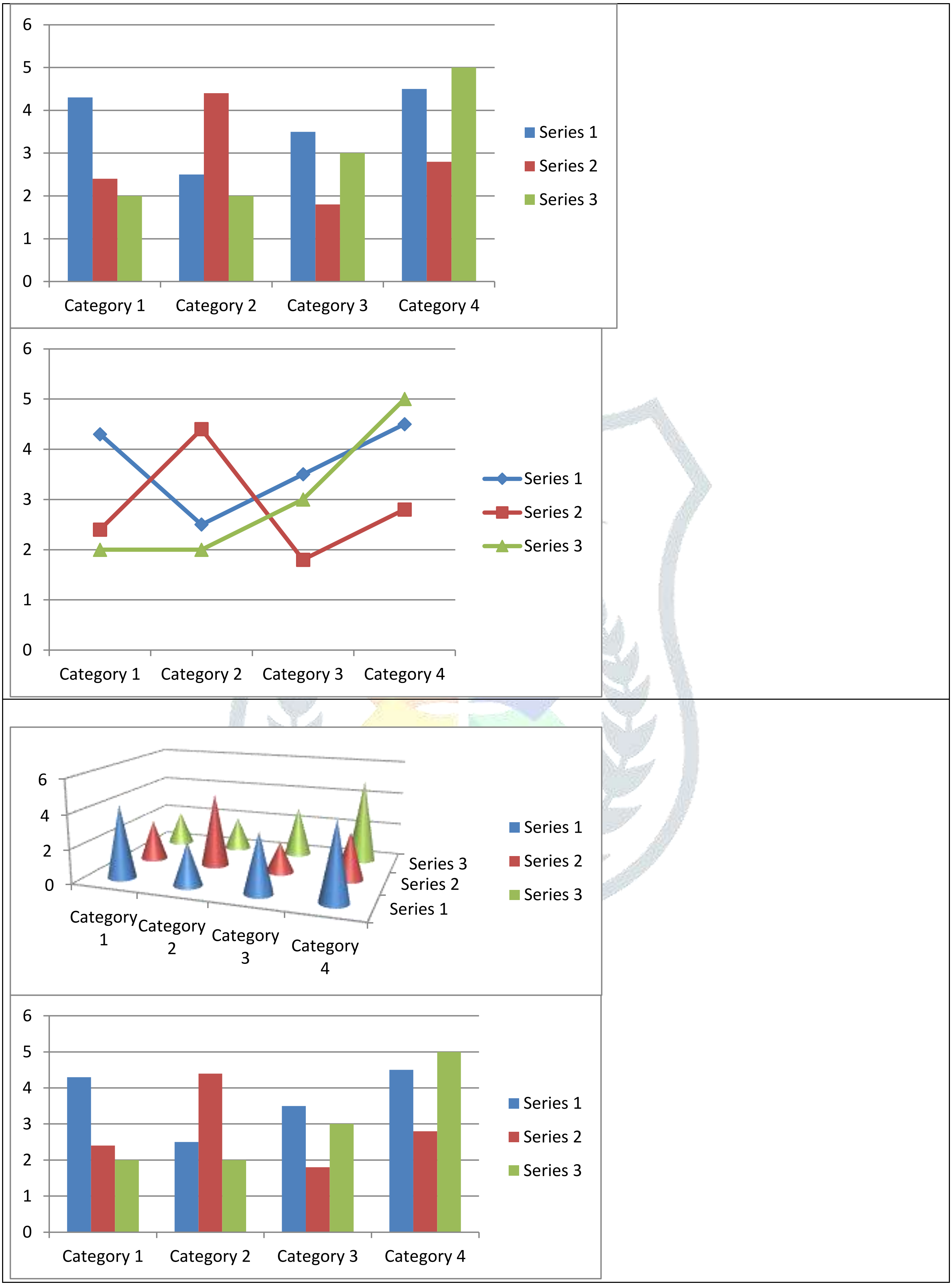


Fig 5.Screen shot for the canopy n 1 cluster.

Cluster#		Full Data	Cluster 0	Cluster 1
1	timedelta	354.5305	357.2445	351.4941
2	n_tokens_title	10.3987	10.1495	10.6776
3	n_tokens_content	546.5147	549.1829	543.5297
4	n_unique_tokens	0.5482	0.53	0.5686
5	n_non_stop_words	0.9965	0.9748	1.0207
6	n_non_stop_unique_tokens	0.6892	0.676	0.7039
7	num_hrefs	10.8837	10.5766	11.2273
8	num_self_hrefs	3.2936	3.3159	3.2688
9	num_imgs4.5441		4.1348	5.0021

10	num_videos1.2499	1.0252	1.5013
11	average_token_length4.5482	4.5824	4.51
12	num_keywords 7.2238	7.2018	7.2483
13	data_channel_is_lifestyle 0.0529	0.0531	0.0527
14	data_channel_is_entertainment 0.178	0.1484	0.2112
15	data_channel_is_bus0.1579	0.1683	0.1462
16	data_channel_is_socmed 0.0586	0.0614	0.0554
17	data_channel_is_tech 0.1853	0.1972	0.1719
18	data_channel_is_world0.2126	0.2517	0.1688
19	kw_min_min 26.1068	26.5886	25.5678
20	kw_max_min 1153.9517	1112.0449	1200.8351
21	kw_avg_min 312.367	307.9767	317.2786
22	kw_min_max 13612.3541	12223.7366	15165.875
23	kw_max_max 752324.0667	750359.1458	754522.3291
24	kw_avg_max 259281.9381	253285.9262	265989.9985
25	kw_min_avg 1117.1466	1062.0118	1178.8289
26	kw_max_avg 5657.2112	5400.942	5943.9132
27	kw_avg_avg3135.8586	3021.4272	3263.8793
28	self_reference_min_shares3998.7554	3928.4182	4077.4454
29	self_reference_max_shares 10329.2127	9848.1021	10867.4569
30	self_reference_avg_sharess 6401.6976	6197.488	6630.1577
31	weekday_is_monday 0.168	0.1703	0.1655
32	weekday_is_tuesday 0.1864	0.1892	0.1833
33	weekday_is_Wednesday 0.1875	0.193	0.1814
34	weekday_is_thursday 0.1833	0.1824	0.1843
35	weekday_is_friday 0.1438	0.1458	0.1416
36	weekday_is_saturday 0.0619	0.0587	0.0654
37	weekday_is_sunday 0.069	0.0606	0.0785
38	is_weekend 0.1309	0.1193	0.1439
39	LDA_00 0.1846	0.1929	0.1753
40	LDA_01 0.1413	0.1224	0.1624
41	LDA_02 0.2163	0.2468	0.1823
42	LDA_03 0.2238	0.1921	0.2592
43	LDA_04 0.234	0.2459	0.2208
44	global_subjectivity 0.4434	0.434	0.4538
45	global_sentiment_polarity 0.1193	0.1157	0.1234
46	global_rate_positive_words 0.0396	0.0372	0.0423
47	global_rate_negative_words0.0166	0.0157	0.0176
48	rate_positive_words 0.6822	0.6825	0.6817
49	rate_negative_words 0.2879	0.2921	0.2833
50	avg_positive_polarity 0.3538	0.3497	0.3584
51	min_positive_polarity 0.0954	0.0953	0.0956
52	max_positive_polarity 0.7567	0.7438	0.7712
53	avg_negative_polarity -0.2595	-0.2529	-0.267
54	min_negative_polarity -0.5219	-0.5127	-0.5323
55	max_negative_polarity -0.1075	-0.106	-0.1091
56	title_subjectivity 0.2824	0.0203	0.5755
57	title_sentiment_polarity 0.0714	0.0031	0.1479
58	abs_title_subjectivity 0.3418	0.4817	0.1854
59	abs_title_sentiment_polarity 0.1561	0.0132	0.3159
60	shares 3395.3802	3220.7258	

Table 1: Canopy clustering Total # of cluster,Cluster 0 and Cluster 1



Graph 1: Canopy clustering Total # of cluster,Cluster 0 and Cluster 1

```
    Time taken to build model (full training data): 2.09 seconds

    === Model and evaluation on training set ===

    Clustered Instances

    0      20032 ( 51%)

    1      19612 ( 49%)
    Log likelihood: -171.76913

    61

=== Cross-validation ===

=== Summary ===

Correlation coefficient                -0.0114
Mean absolute error                  3205.4948
Root mean squared error                11626.9803
Relative absolute error               100      %
Root relative squared error             100      %
Total Number of Instances               39644
```

Conclusion

In the research paper classification and clustering algorithms are evaluated using the big data set.The preprocessing is done using unsupervised learning method AddExpression-E 0.0 expression algorithm. We have implemented two classifiers using ZeroR and RepTree classification.We have successfully implemented two clustering algorithms such as KMeans clustering and Canopy cluster algorithm. Their performances are recorded and compared. The ZeroR gives the best result with less time complexity of 0.05 seconds and canopy clustering took 2.09 seconds. In the future enhancement other classification algorithms can be eevaluated to improve the accuracy and time complexity.

References

1. http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/REPTree.html
2. https://archive.ics.uci.edu/ml/index.php
3. Hensinger, Elena, Ilias Flaounas, and Nello Cristianini. "Modelling and predicting news popularity." Pattern Anal- ysis and Applications 16.4 (2013): 623-635.
4. K. Fernandes, P. Vinagre and P. Cortez. A Proactive In- telligent Decision Support System for Predicting the Popularity of Online News. *Proceedings of the 17th EPIA 2015 Portuguese Conference on Artificial Intelligence*, September, Coimbra, Portugal.
5. Predicting the Popularity of Social News Posts." 2013 cs229 projects. Joe Maguire Scott Michelson

6.Tatar, Alexandru, et al. "Predicting the popularity of online articles based on user comments." Proceedings of the International Conference on Web Intelligence, Mining and Semantics. ACM, 2011.

7. “Predicting and Evaluating the Popularity of Online News”,He Ren and Quan Yang, pp.1-5