Credibility Identification of Online News Portal Using Website Traffic Metrics

Farhad Rahman

ID: 2015-2-60-044

Md. Ashfak Ragib

ID: 2015-2-60-086

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering

Department of Computer Science and Engineering
East West University
Dhaka-1212, Bangladesh

December, 2019

Declaration

We, hereby, declare that the work presented in this thesis is the outcome of the investigation performed by us under the supervision of Tanni Mittra, Senior Lecturer, Department of Computer Science and engineering, East West University. We also declare that no part of this thesis has been or is being submitted elsewhere for the award of any degree or diploma.

Countersigned	Signature
………………………	………………………
(Tanni Mittra)	**(Farhad Rahman)**
Supervisor	**(2015-2-60-044)**
	Signature
	………………………
	(**Md. Ashfak Ragib)**
	(2015-2-60-086)

Abstract

Online news portal sites are increasing progressively as like meshwork all over the world. The propagation of misrepresenting information such as social media feeds, news blogs, and online newspapers have made it challenging to distinguish reliable news sources, thus enhancing the need for computational tools able to provide insights into the trustworthiness of online content. It’s a much difficult task to determine the authenticity of the newspaper article content directly rather than knowing the actual source reliability of the article. To distinguish reliable and unreliable media sources, an approach is proposed in this paper to scale online news portal reliability using website metrics data that has been collected through the Alexa website traffic statistics tool. Initially, all Bengali online news portal websites were listed and a dataset has been created containing all relevant website metrics information of all those websites. Using domain knowledge and context analysis vital features have been extracted for scaling reliability of those websites using an unsupervised learning algorithm. In this context, the k-means algorithm is been used for making several clusters of the unlabeled dataset. Each cluster got a label depending on the metrics information correlation in terms of the real-world scenario. Comparing the experimental result with the theoretical knowledge, the proposed approach satisfied the research intention.

Acknowledgments

As it is true for everyone, we have also arrived at this point of achieving a goal in our life through various interactions with and help from other people. However, written words are often elusive and harbor diverse interpretations even in one's mother language. Therefore, we would not like to make efforts to find best words to express our thankfulness other than simply listing those people who have contributed to this thesis itself in an essential way. This work was carried out in the Department of Computer Science and Engineering at East West University, Bangladesh.

First of all, we would like to express our deepest gratitude to the almighty for His blessings on us. Next, our special thanks go to our supervisor, "Tanni Mittra", who gave us this opportunity, initiated us into the field of "An Approach To Scale Online News Portal Reliability Analyzing Website Metrics Data", and without whom this work would not have been possible. His encouragements, visionaries and thoughtful comments and suggestions, unforgettable support at every stage of my/our BS.c study were simply appreciating and essential. His ability to muddle us enough to finally answer our own question correctly is something valuable what we have learned and we would try to emulate if ever we get the opportunity.

We would like to thank "Sarwar Sajol" for his excellent collaboration during performance evaluation studies; "Touhidul Islam Shoheb" for his overall support; "Kazi Khaled Saif Ullah" for her helpful suggestions in solving tricky technical problems. Last but

not the least, we would like to thank our parents for their unending support, encouragement and prayers.

There are numerous other people too who have shown me their constant support and friendship in various ways, directly or indirectly related to our academic life. we will remember them in our heart and hope to find a more appropriate place to acknowledge them in the future.

Farhad Rahman
December, 2019

Md. Ashfak Ragib
December, 2019

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recent advanced in modern age of technology, heterogeneous information are available through online resources. With the growing popularity of the WWW, Websites are playing a significant role in conveying knowledge and information to the users [1]. Internet is one of the best way of becoming the universal communication network for gathering and sharing all kinds of information, from the simple transfer of binary computer data to the transmission of voice, video, or interactive information in real-time [2].

Finding the legitimate source of online newspapers, data collection and analysis, we use "Web Alexa" which is owned by "Amazon.com".By entering the website's domain and Alexa will disclose the ranking of that website based on a combined measure of unique visitors and page views.

1.1 Overview & Motivation

It's challenging to identify the reliability and unreliability of the website with various kinds of miscellaneous information. For example, online newspapers are rapidly growing because of its flexibility and availability to the audience. As a result, online newspapers are become very reliable entities and plays a vital role in this current era. Leading misinformation, peoples are beyond the truth of real news and some group take the opportunity by spreading the false news.

1.2 Problem Statement

Recently it is hard to believe that, using the Internet, social networking sites (SNS) and click through rate (CTR) based policies have made it possible to create hoaxes, large scale published and propagate it to a extensive number of audience with a higher speed than ever [3].

It is possible to detect the suspicious source through source detection system. In this paper, we have proposed a methodology to detect online newspapers source and predict them whether this source is reliable or not.

Forged news, which refers to false dictate advertised under the masquerade of being reliable news [4]. There has many system to detect fake news or reliability checking system [5], but in this paper we will predict the reliability of news sources.

1.3 Objectives

The aim of this study is to predict online newspaper sources' reliability or unreliability using unsupervised learning. Because numerous online news websites are spreading rumor nowadays. But the general audience doesn’t know the origin or the root of this news. Eventually, many social sites are rumoring forge news instead of real news. Our main three

Objectives are-

- To create a model by web trafficking and analysis of online Bangla newspaper sources.
- To find out the reliability or unreliability of the sources by the supervised and unsupervised learning system.
- To compare and the result analysis of supervised and unsupervised learning systems.

1.4 Contribution

Within this paper, at first, we have created our dataset by website traffic of various online newspapers from ”Alexa Internet”, provided by ”Amazon.com”.For the prediction of reliability or unreliability, we have used 5 main features including Sites, BD Rank, Pageviews/User, Bounce Rate, Time On Site(minutes), Search, Social, Link, Direct link, Referral Sites. Since there is no class label in our dataset and based on these features, we will use the clustering technique to develop the prediction system of online newspaper sources reliability or unreliability.

2 Find out the important features after doing the web trafficking of online newspapers.

3 After finding the important features, we develop the clustering model.

4 At first, we will cluster the whole dataset with raw data and we’ll get a cluster label.

5 Then we will cluster individually different types of attributes and again we’ll get different cluster labels.

6 Getting all cluster labels, we will ensemble the labels and compare them with our raw data cluster label.

7 Comparing both labels, we can predict the reliability of these Sites sources.

8 Finding from the compare of clusters labels, we’ll use the classification system to get better accuracy and gain the prediction correctly.

1.5 Thesis Organization

We organized the rest of the paper as follows. In section 2, we discuss the different existing approaches of web trafficking, predict the website source reliability through clustering technique, classify them, etc. to introduce the priority of online news websites sources and its effects in daily life. In section 3, we propose our methodology. We discuss how we collect data, clustering, classify, measure, and predict the news site's reliability. In section 4, we discuss the materials that we have used in this experiment. We also discuss different environments, tools, data reduction techniques, etc. Section 5 describes our dataset, experiment with the dataset, result analysis, and discussion, while section 6 concludes.

Chapter 2

Literature Review

In this section, we have discussed about the existing approach to analyze the website metrics data and brief description of website evaluation, Alexa internet, hoax news and website spoofing. Website evaluation provides useful information for users to estimate sites validation and popularity. So far, a number studies using web metrics methods have been done by various authors on different websites. Here is an attempt is made portray some of the website evaluation studies using Alexa Internet as tool for evaluation.

2.1 Existing Alexa Technique

Since early 1996, Alexa informs that its crawler has been crawling the Web. It has "gathered 4.5 billion pages from over 16 million sites" (Alexa Internet, 2011a). The users who use Alexa toolbar are the primary source of Alexa traffic data (traffic ranking and page views etc.). To become a toolbar user, anyone can download and use Alexa toolbar. From its toolbar users, Alexa collects Web traffic data, "Alexa users have downloaded millions of Toolbars" (Alexa Internet, 2011b).For academic impact measure and business performance measure, Vaughan (2008) explained Alexa traffic data and reported studies that used the traffic data [6].In earlier Alexa has used Web metrics [7] [8]studies as a valuable data source, especially its traffic data [9].

2.2 Web Metrics system using Alexa Internet

Different authors have used web metrics system on various websites because of sites validation and demand for users. As website evaluation models, Shen et al. (2006) evaluated 15 university library web sites using Alexa Internet as a tool based on traffic rank, connectivity, visits, speed, freshness, and pages viewed. Using Alexa Internet, Indian newspaper websites evaluated by Bhat (2013). Clustering and several criteria gathered from the Alexa search engine used to evaluate Greek newspaper by Kanellopoulos and Kotsiantis (2012) [10].

2.3 Hoax News And Website Spoofing

Website spoofing use for creating a website as a hoax and the prime intention is to mislead or confusing the readers. It is structured to make visitors believe they are visiting trusted sources like BBC News or The New York Times. Fictitious articles purposely shaped to hype readers, usually with the target of benefiting through clickbait [11].

In this modern age, people are living with the internet and connected by various social sites and media like (facebook, tweeter, Instagram, snap chat, etc.). Eventually, a network always surrounds them and getting different kinds of news from their social sites. With considering these news, several sources are authentic, and several are not. A piece of false news can be a great disaster for society. The latest example of rumour related to “600 Murders Take Place in Chicago during the second weekend of August 2018” presents fear and anxiety about such a large number of violence in the city”. Source detection of rumour in the social network is the research direction [12].

2.3 Evaluating News Website

Using Alexa databank, the primary purpose is to evaluate Indian newspaper websites. Audit Bureau of Circulations has compiled the list of top newspapers in India by daily circulation was used for selecting newspapers. The list carried 28 newspapers in various languages. Out of these, 26 were available online. For evaluation, these 26 newspapers are taken in the present study. In Alexa databank, each newspaper web site had searched, and associated data including traffic rank, pages viewed, speed, links, bounce percentage, time on site, search percentage, and Indian/foreign users were collected. Collected data were analyzed and tabulated to reveal findings following the desired objectives [13].

Chapter 3

Proposed Methodology

The present study has been done by using webometric methods with the help of Alxa databank. In this research we selected eleven features or attributes – i.e. Site name,pages viewed, social, links, bounce percentage, time on site, search percentage, global rank,country rank(BD), direct link, referral sites. – In order to analyze online newspaper websites credibility.

3.1 Data Collection Method

The first step inflow of the work is collecting data from Alexa by using website traffic statistic tool. The first task is listing available online Bangladeshi newspapers. After completing the first task, the next task is to search each online newspapers and collect data using eleven attributes. Then collected data has been set up to excel datasheet according to their proper attribute. . After organizing the dataset, the machine learning algorithm has implemented. Specifically unsupervised learning method. All the properties of data collection depending on tools and feature based on data collection description are given below:

3.1.1 Alexa Web Traffic Analysis

Alexa traffic ranks are based on two sets of data. The total number of unique visitors as well as page views. In other words, top-ranked sites are not only able to attract new users but also have quality content that is able to keep people on site.Rank on Alexa.com is calculated based on global data that was accrued over the last three months. The traffic rank is updated daily, which makes it pretty accurate.

If user open the same page several times during a day, the system will count it as a single view. In other words, user will not be able to click on the same page several times and thus increase view count. It is a great way to prevent a user from artificially boosting stats. However, visitors can directly improve online rankings if he visits different pages.This service can be used free of charge through Alexa’s website.

Users can also install the Alexa toolbar for your Google Chrome browser allowing to check domain rankings at any time. Once installed, it also provides some other features:

- Related Links – You can use it to find websites that are similar to the one that you’re currently visiting
- Wayback – Check how different blogs looked in the past
- Search Analytics – Learn more about popular keywords which can help you increase reach

3.1.2 Features Representation

There has some features to find out and scaling the reliability. These features or attributes are clollected from Alexa by using website traffic statistics tool. From these tool we adopted eleven main features which will be fulfil our research intention. Few attributes description are given below:

1. **Page Views Per User:**

   In the world of online publishers, the number and quality of website visitors received are critical. Without website visitors, there can be no ad sales and no business. It makes understanding and optimizing metrics like pageviews all the more critical.

   Since Google Analytics is one of the most popular website traffic tracking and analytical tools you'll see it mentioned many times in this post. Let's take a closer look at what pageviews are and how they work.A simple definition for a page view can relate to a user viewing a web page. A user can also browse more than one page per website visit.

   Google Analytics describes a pageview as the following:

   "A view of a page on your site that is being tracked by the Analytics tracking code." Also when a user visits more than one page, they are counted as additional pageviews.

2. **Bounce Rate:**

   Bounce rate is a metric that measures the percentage of people who land on your website and do completely nothing on the page they entered. So they don't click on a menu item, a 'read more' link or any other internal links on the page. This means that the Google Analytics server doesn't receive a trigger from the visitor. A user bounces when there has been no engagement with the landing page and the visit ends with a single-page visit. You can use bounce rate as a metric that indicates the quality of a webpage and/or the "quality" of your audience. By the quality of your audience I mean whether the audience fits the purpose of your site.

3. **Time On Site:**

   In Web analytics, including Google Analytics, average time on site is a type of visitor report that provides data on the amount of time (in minutes or seconds) visitors have spent on your website. When viewing the time on site report in your Web analytics program it is important to remember that the results can be misleading because in some cases the visitor may have been interacting with your pages and site content or they could have left the browser window open and were not actually viewing your page.

4. **Referral Sites:**

   A referral website is an Internet address or hostname used to get a visitor to another site. A visitor has clicked a hyperlink on the referral website, which leads to the website where he is located now. The referral website is thus the source for the traffic on the current page. When a website is viewed, the visitor's browser transmits the name of the requested website and also the origin of the referral link that directed the visitor to it. Referral websites are important data in web analysis in order to assign traffic to the different sources and to find out where the visitors of a website came from. The terms referrer, referrals or referral traffic are also common in addition to referral website.

3.2 Clustering Method

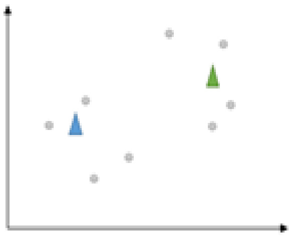
Clustering refers to a broad set of techniques for finding subgroups or clusters in a dataset. This helps us partition observations into distinct groups so that each group contains observations that are similar to each other. There are many clustering methods, but we will focus on **k-means clustering** and **hierarchical clustering**.

3.2.1 K-Means Clustering

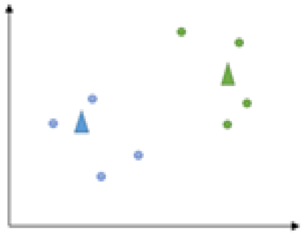
The first step of this algorithm is creating, among unlabeled observations, *c* new observations, randomly located, called 'centroids'. The number of centroids will be representative of the number of output classes (which, remember, we do not know). Now, an iterative process will start, made of two steps:

- First, for each centroid, the algorithm finds the nearest points (in terms of distance that is usually computed as Euclidean distance) to that centroid, and assigns them to its category;
- Second, for each category (represented by one centroid), the algorithm computes the average of all the points which has been attributed to that class. The output of this computation will be the new centroid for that class.

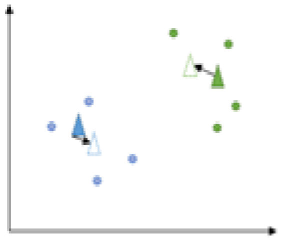
Every time the process is reiterated, some observations, initially classified together with one centroid, might be redirected to another one. Furthermore, after several reiterations, the change in centroids' location should be less and less important since the initial random centroids are converging to the real ones. This process ends when there is no more change in centroids' position.



1. Two centroids are randomly located among data



2. The algorithm finds the nearest observations for each centroid, and then attributes them to each centroid's class



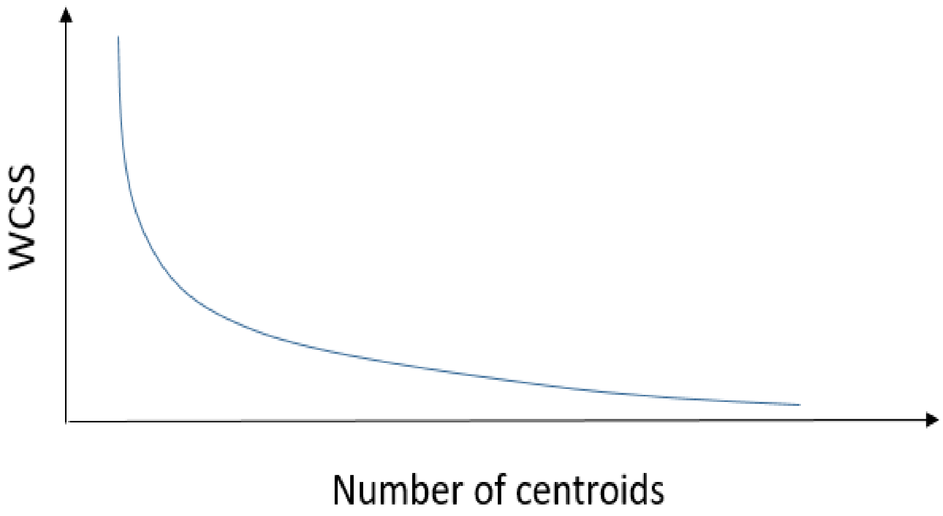
3. New centroids, resulting from the average of each class observations, are computed, and then the process is repeated

There are many methods that could be employed for this task. However, in this paper, we are going to explain and use the so 'Elbow method'. The idea is that what we would like to observe within our clusters is a low level of variation, which is measured with the within-cluster sum of squares (WCSS):

$$WCSS = \sum_{xi \in c} (x_i - \bar{x})^2$$

And it is intuitive to understand that, the higher the number of centroids, the lower the WCSS. In particular, if we have as many centroids as the number of our observations, each WCSS will be equal to zero. However, if we remember the law of parsimony, we know that setting the highest number possible of centroids would be inconsistent.

The idea is picking that number of centroids after which the reduction in WCSS is irrelevant. The relation I've just described can be represented with the following graph:



The idea is that, if the plot is an arm, the elbow of the arm is the optimal number of centroids.

3.2.2 Hierarchical Clustering

This algorithm can use two different techniques:

- Agglomerative
- Divisive

Those latter are based on the same ground idea, yet work in the opposite way: being *K* the number of clusters (which can be set exactly like in K-means) and *n* the number of data points, with *n*>*K*, agglomerative HC starts from *n* clusters, then aggregates data until it obtains K clusters; divisive HC. On the other hand, starts from just one cluster and then splits it depending again on similarities until it obtains *K* clusters.

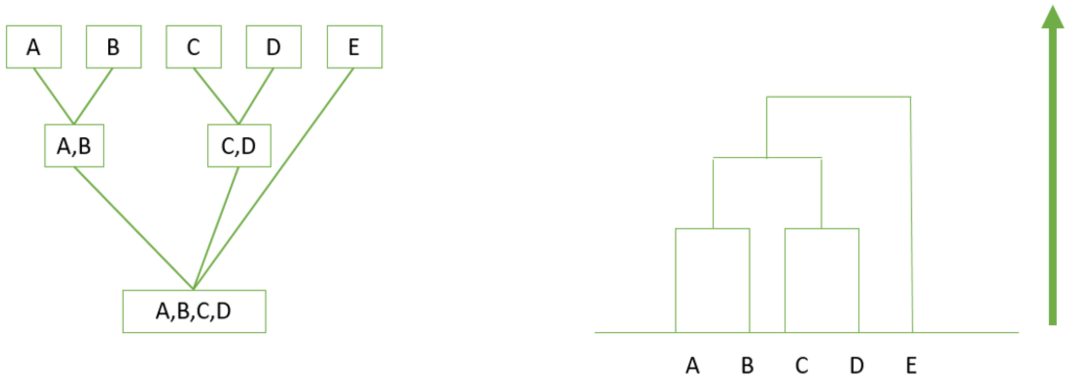


Figure 1: Agglomerative Hierarchical Clustering

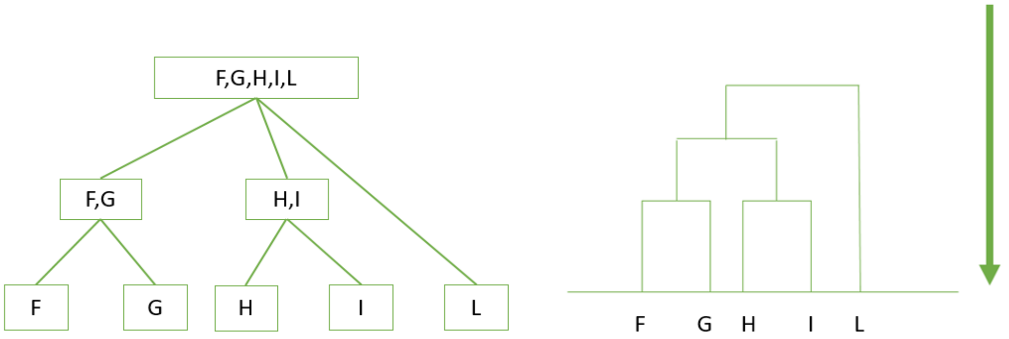


Figure 2: Divisive Hierarchical Clustering

As anticipated, the key element of discrimination here is similarity among data points. In mathematical terms, similarity mainly refers to distance, and it can be computed with different approaches. Here, I will propose three of them:

- **MIN:** it states that, given two clusters C1 and C2, the similarity between them is equal to the minimum of similarity (translated: distance) between point a and b, such that a belongs to C1 and b belongs to C2.

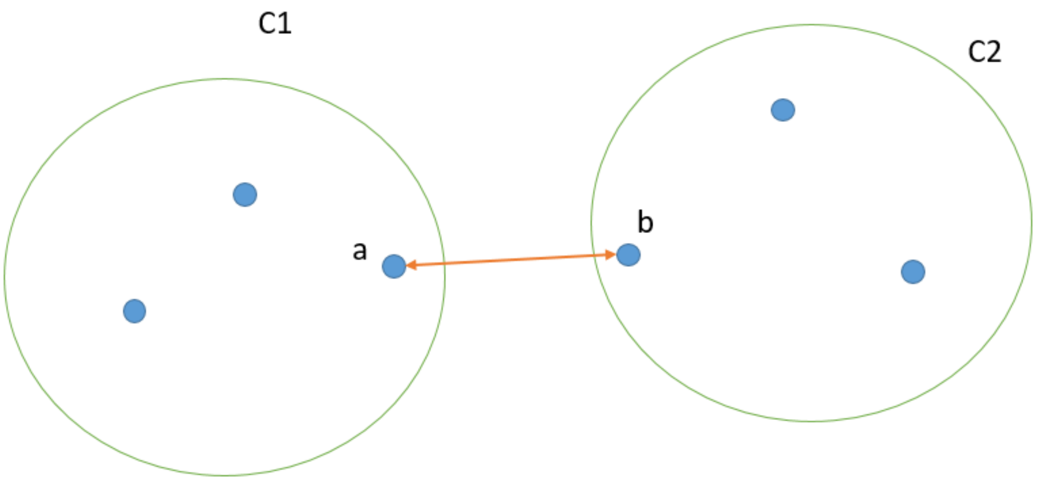


Figure 3: MIN

- **MAX:** it states that, given two clusters C1 and C2, the similarity between them is equal to the maximum of similarity between point a and b, such that a belongs to C1 and b belongs to C2.

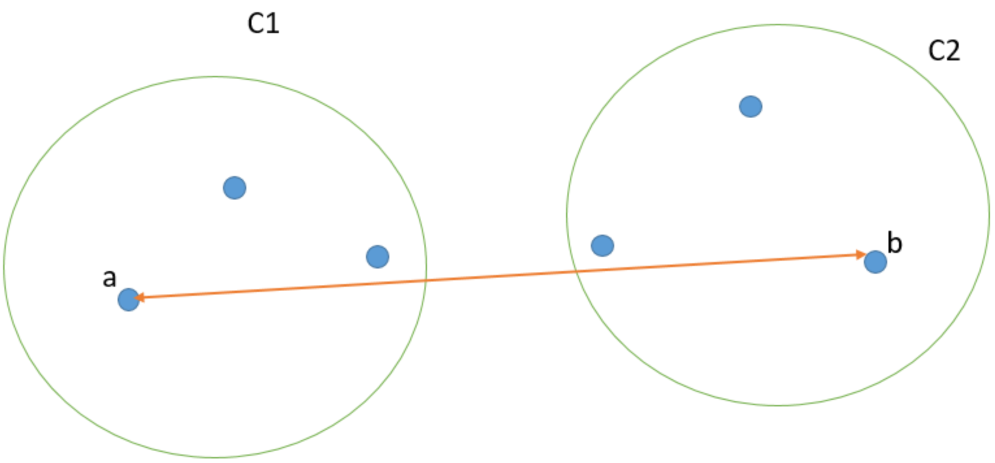


Figure 4: MAX

- **Average:** it takes all the pairs of points, compute their similarities and then calculate the average of the similarities. That latter is the similarity between the clusters C1 and C2.

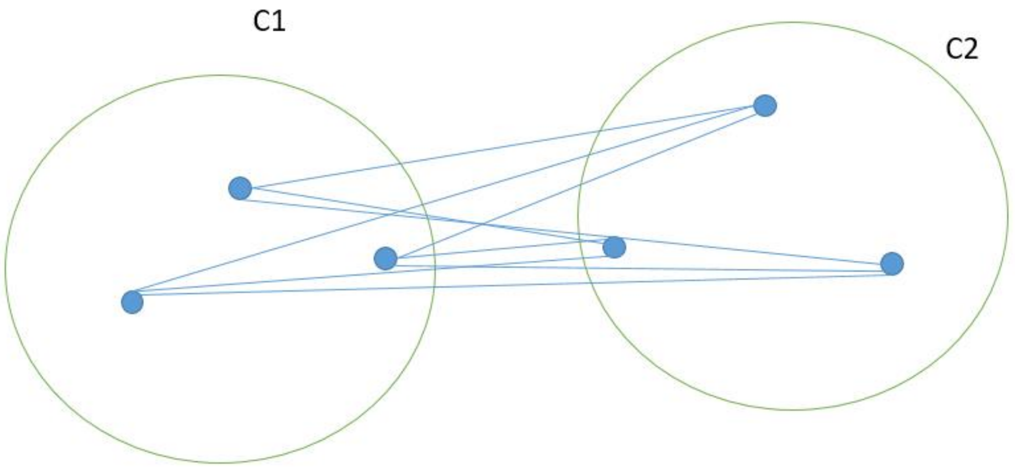


Figure 5: Average

So, both the algorithms look for similarities among data and both use the same approaches to decide the number of clusters.

Chapter 4

Experiment and Result Analysis

We have implemented the proposed approach in Jupyter Notebook, Jupyter is a non-profit, open-source project, born out of the IPython Project in 2014 as it evolved to support interactive data science and scientific computing across all programming languages. We have applied our proposed method on our own made dataset which has been created using Alexa internet traffic analysis tool. The dataset contains all available Bengali online newspaper websites metrics data. All the experiments are carried out on an Intel Core2 Duo CPU E8400 @3.00GHz desktop machine with 4 GB DDR3 memory under Windows 10 Professional operating system and python-3 compiler. In this section, we will discuss the result attained in our experiments on various classification algorithm.

4.1 Experiment With Dataset-1

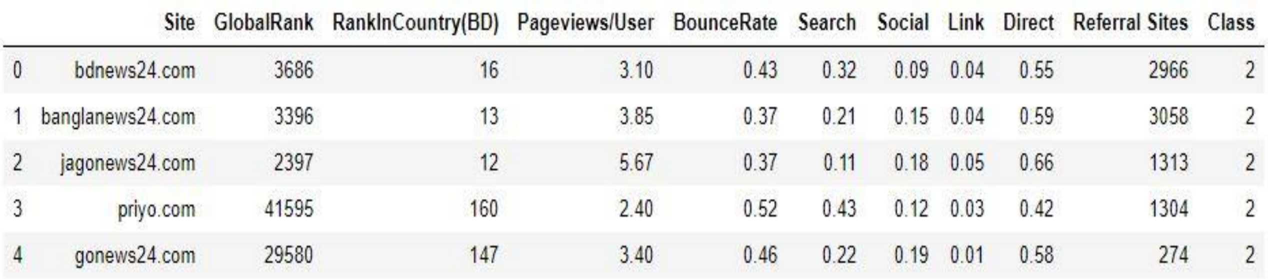
Dataset-1 consist of after clustering on online news portal webmetric dataset and then make a cluster label column based on apllying clustering method. This subsection contains detail description of evaluation datasets with experimental methodology and performance analysis. The performance of the proposed method is measured based on classification algorithm evaluation metrics.

4.1.1 Evaluation Dataset

In our experiment, as mentioned above, we have used the dataset that has been produced with using Alexa internet traffic analysis tool where the dataset was unlabeled data. After feature extraction and then applying the k-means unsupervised learning algorithm on the dataset in straight forward manner we got the clusters, that helped us to labeled the dataset. After applying the clustering algorithm we got 4 clusters, where the cluster number was obtained using elbow graph analysis. From the domain knowledge and analyzing correlation between every website attributes we labelled those clusters as Highly Reliable, Reliable, Semi-Reliable and Unreliable. But for computational purpose we used numerical class label which is defined from 1 to 4. After adding the class label our dataset contains 11 attributes and 55 rows.

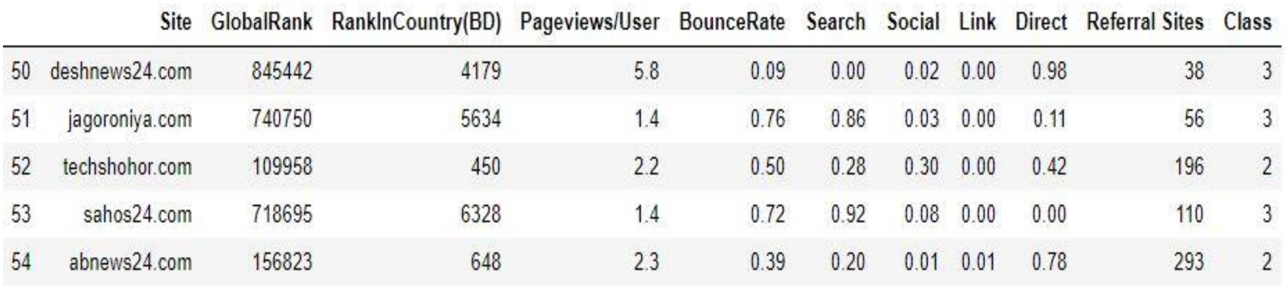
The attributes of the dataset are Site,

Global Rank, Rank In Country (BD), Pageviews/ User, Bounce Rate, Search, Social, Link, Referral Sites, Class.



	Site	GlobalRank	RankInCountry(BD)	Pageviews/User	BounceRate	Search	Social	Link	Direct	Referral Sites	Class
0	bdnews24.com	3686	16	3.10	0.43	0.32	0.09	0.04	0.55	2966	2
1	banglanews24.com	3396	13	3.85	0.37	0.21	0.15	0.04	0.59	3058	2
2	jagonews24.com	2397	12	5.67	0.37	0.11	0.18	0.05	0.66	1313	2
3	priyo.com	41595	160	2.40	0.52	0.43	0.12	0.03	0.42	1304	2
4	gonews24.com	29580	147	3.40	0.46	0.22	0.19	0.01	0.58	274	2

Figure 6: Dataset-1 Head



	Site	GlobalRank	RankInCountry(BD)	Pageviews/User	BounceRate	Search	Social	Link	Direct	Referral Sites	Class
50	deshnews24.com	845442	4179	5.8	0.09	0.00	0.02	0.00	0.98	38	3
51	jagoroniya.com	740750	5634	1.4	0.76	0.86	0.03	0.00	0.11	56	3
52	techshohor.com	109958	450	2.2	0.50	0.28	0.30	0.00	0.42	196	2
53	sahos24.com	718695	6328	1.4	0.72	0.92	0.08	0.00	0.00	110	3
54	abnews24.com	156823	648	2.3	0.39	0.20	0.01	0.01	0.78	293	2

Figure 7: Datase-1 Tail

4.1.2 Experimental Methodology

We have experimented three types of classification algorithm on our dataset, where the classification algorithms are Decision Tree, Random Forest and K-Nearest Neighbor classifier. While applying those algorithms on our dataset we tried to tune the parameters of the classification algorithm to obtain better accuracy. That’s why we got different accuracy result in different stages of the experiment and we closed this tuning of parameters when the highest accuracy was obtained.

Those classification algorithms are been used mainly to evaluate the dataset labelling or clustering quality. If the dataset labelling quality is good then the classification algorithm would give a good result in terms of accuracy measurement that defines the strong correlation between attributes and the data are been clustered properly.

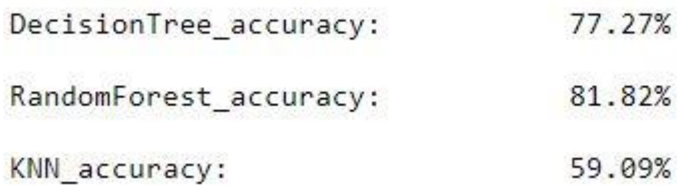
For classification algorithm accuracy measurement there are various types of measurement metrics. Accuracy, Precision, Recall, F1- score etc. we have used those measurement metrics to evaluate the algorithms performance and to perform comparison between those algorithms to identify the better result giving classification algorithm for this dataset.

4.1.2 Experimental Result

Experimental results of our proposed method as elaborated below, can be judged based on the aspects which have been mentioned previously.

Accuracy Percentage

After applying the above mentioned algorithms on the dataset we have found the accuracy result of those algorithms for the dataset and plotted the result in bar chart. Here, the mentioned result of the accuracy are highest after tuning the algorithm parameters value.



```
DecisionTree_accuracy:          77.27%

RandomForest_accuracy:          81.82%

KNN_accuracy:                   59.09%
```

Figure 8: Dataset-1 Model Accuracy

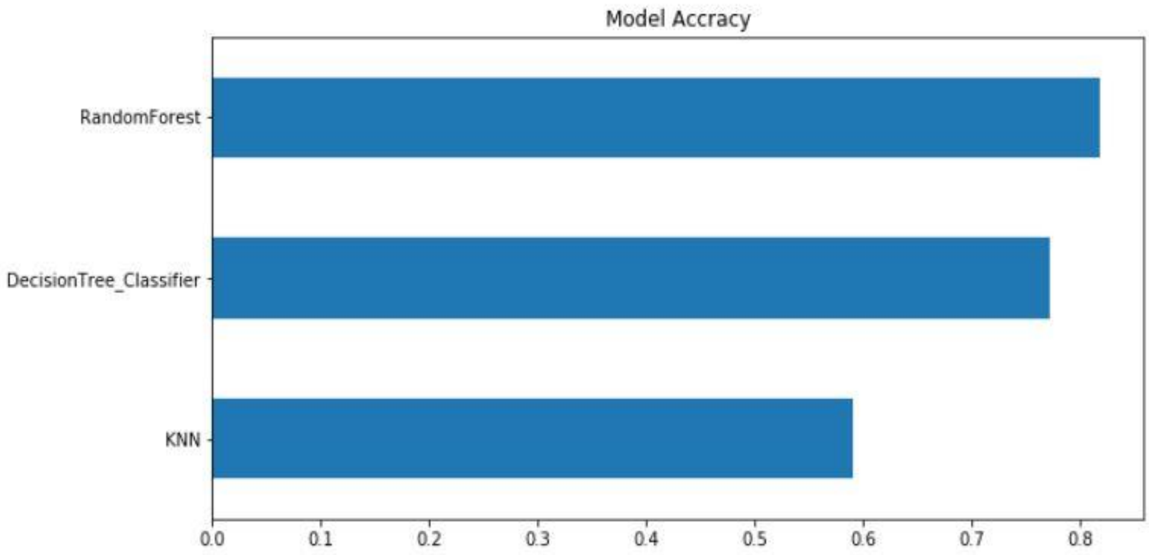
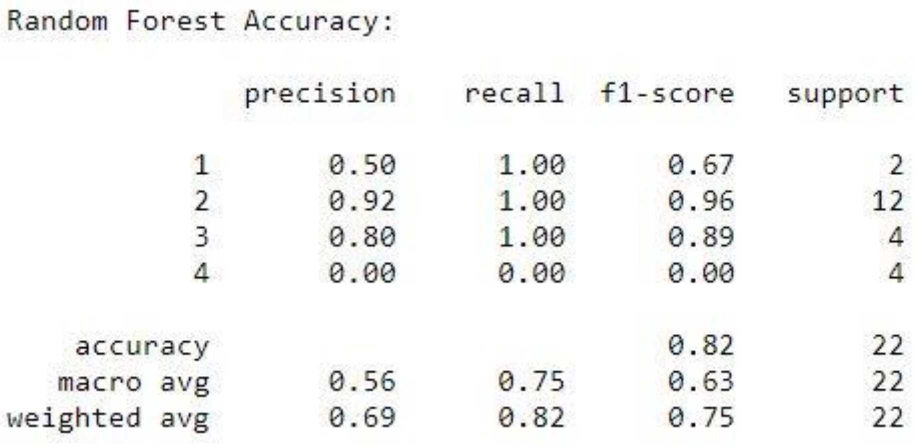


Figure 9: Dataset-1 Model Accuracy Chart

After finding the accuracy percentage of the models we have investigated about precision, recall and f-1 score of the models for getting better insight about the accuracy. Result about the investigation for individual models are mentioned below.



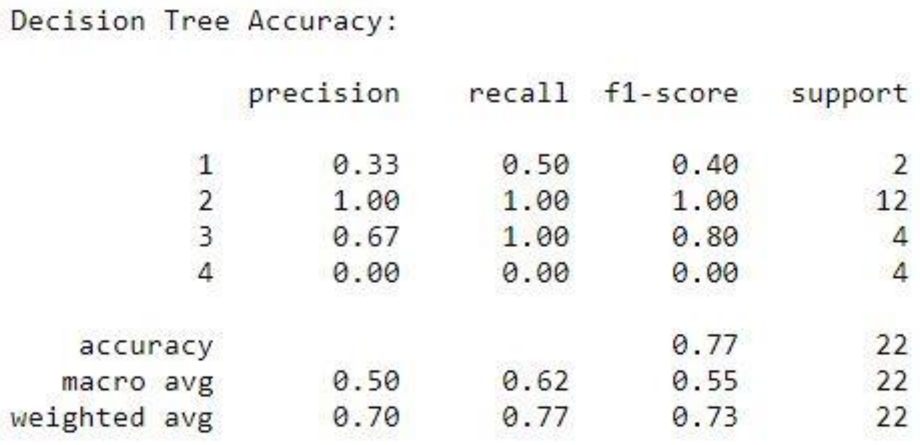
```
Random Forest Accuracy:

              precision    recall  f1-score   support

           1       0.50      1.00      0.67         2
           2       0.92      1.00      0.96        12
           3       0.80      1.00      0.89         4
           4       0.00      0.00      0.00         4

    accuracy                           0.82        22
   macro avg       0.56      0.75      0.63        22
weighted avg       0.69      0.82      0.75        22
```

Figure 10: Dataset-1 Random Forest classification report



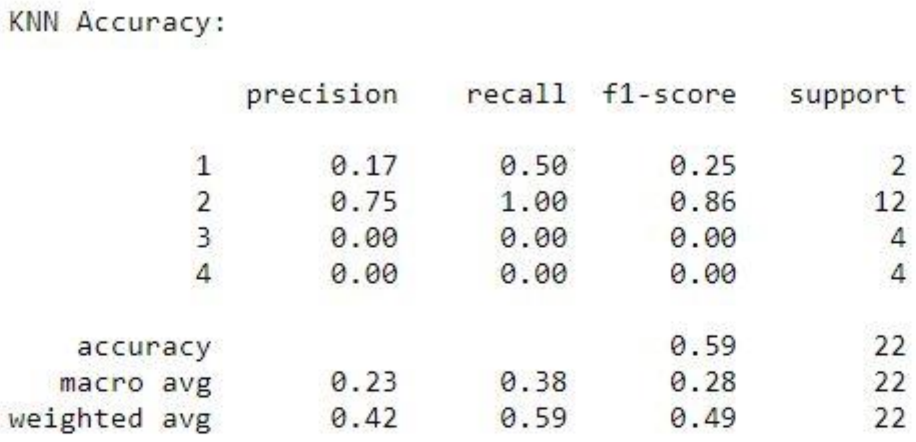
```
Decision Tree Accuracy:

              precision    recall  f1-score   support

           1       0.33      0.50      0.40         2
           2       1.00      1.00      1.00        12
           3       0.67      1.00      0.80         4
           4       0.00      0.00      0.00         4

    accuracy                           0.77        22
   macro avg       0.50      0.62      0.55        22
weighted avg       0.70      0.77      0.73        22
```

Figure 11: Dataset-1 Decision Tree classification report



```
KNN Accuracy:

              precision    recall  f1-score   support

           1       0.17      0.50      0.25         2
           2       0.75      1.00      0.86        12
           3       0.00      0.00      0.00         4
           4       0.00      0.00      0.00         4

    accuracy                           0.59        22
   macro avg       0.23      0.38      0.28        22
weighted avg       0.42      0.59      0.49        22
```

Figure 12: Dataset-1 KNN classification report

4.2 Experiment With Dataset-2

Dataset-2 consist of after partioned cluster on online news portal webmetric dataset and then maerge the partitioned cluster and make a cluster label column based on apllying clustering method. We also experimented our dataset-2 with the same algorithms that have been used in the previous experiment section. In this section we have followed the same process that included evaluating dataset with experimental methodology and performance analysis.

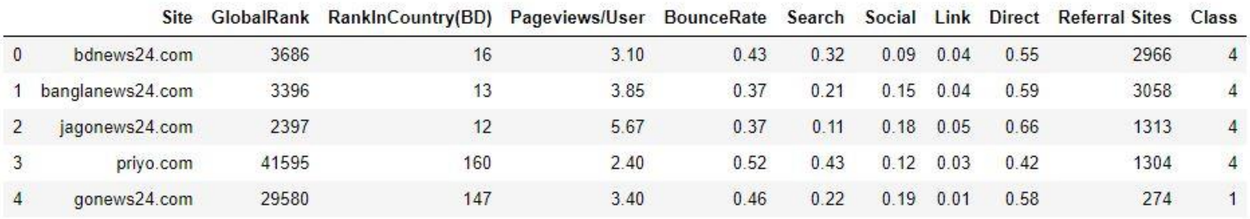
4.2.1 Evaluation Dataset

In our experiment, as mentioned above, we have used the dataset that has been produced with using Alexa internet traffic analysis tool where the dataset was unlabeled data. After feature extraction and then applying the k-means unsupervised learning algorithm on the dataset in straight forward manner we got the clusters, that helped us to labeled the dataset. After applying the clustering algorithm we got 4 clusters, where the cluster number was obtained using elbow graph analysis. From the domain knowledge and analyzing correlation between every website attributes we labelled those clusters as Highly Reliable, Reliable, Semi-Reliable and Unreliable. But for computational purpose we used numerical class label which is defined from 1 to 4. After adding the class label our dataset contains 11 attributes and 55 rows.

The attributes of the dataset are Site,

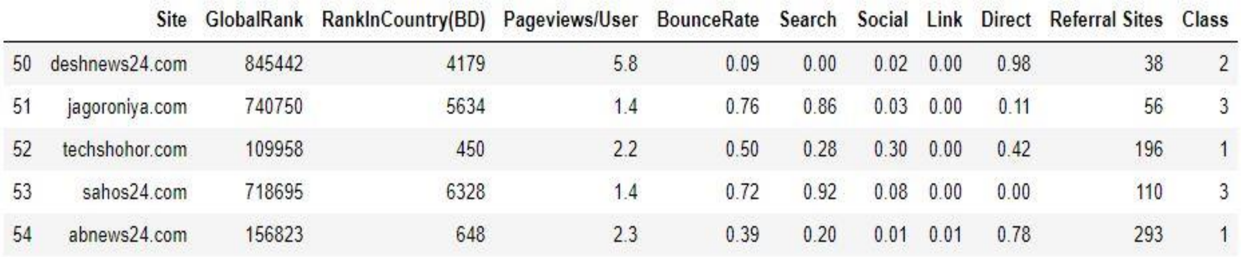
Global Rank, Rank In Country (BD), Pageviews/ User, Bounce Rate,

Search, Social, Link, Referral Sites, Class.



	Site	GlobalRank	RankInCountry(BD)	Pageviews/User	BounceRate	Search	Social	Link	Direct	Referral Sites	Class
0	bdnews24.com	3686	16	3.10	0.43	0.32	0.09	0.04	0.55	2966	4
1	banglanews24.com	3396	13	3.85	0.37	0.21	0.15	0.04	0.59	3058	4
2	jagonews24.com	2397	12	5.67	0.37	0.11	0.18	0.05	0.66	1313	4
3	priyo.com	41595	160	2.40	0.52	0.43	0.12	0.03	0.42	1304	4
4	gonews24.com	29580	147	3.40	0.46	0.22	0.19	0.01	0.58	274	1

Figure 13: Dataset-2 Head



	Site	GlobalRank	RankInCountry(BD)	Pageviews/User	BounceRate	Search	Social	Link	Direct	Referral Sites	Class
50	deshnews24.com	845442	4179	5.8	0.09	0.00	0.02	0.00	0.98	38	2
51	jagoroniya.com	740750	5634	1.4	0.76	0.86	0.03	0.00	0.11	56	3
52	techshohor.com	109958	450	2.2	0.50	0.28	0.30	0.00	0.42	196	1
53	sahos24.com	718695	6328	1.4	0.72	0.92	0.08	0.00	0.00	110	3
54	abnews24.com	156823	648	2.3	0.39	0.20	0.01	0.01	0.78	293	1

Figure 14: Dataset-2 Tail

4.2.2 Experimental Methodology

We have experimented three types of classification algorithm on our dataset, where the classification algorithms are Decision Tree, Random Forest and K-Nearest Neighbor classifier. While applying those algorithms on our dataset we tried to tune the parameters of the classification algorithm to obtain better accuracy. That’s why we got different accuracy result in different stages of the experiment and we closed this tuning of parameters when the highest accuracy was obtained.

Those classification algorithms are been used mainly to evaluate the dataset labelling or clustering quality. If the dataset labelling quality is good then the classification algorithm would give a good result in terms of accuracy measurement that defines the strong correlation between attributes and the data are been clustered properly.

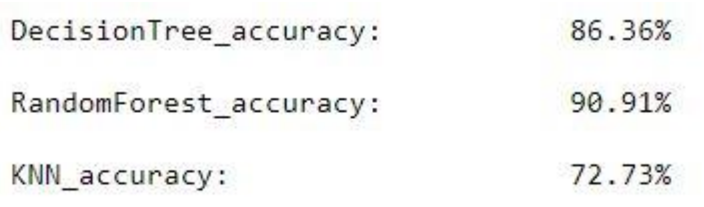
For classification algorithm accuracy measurement there are various types of measurement metrics. Accuracy, Precision, Recall, F1- score etc. we have used those measurement metrics to evaluate the algorithms performance and to perform comparison between those algorithms to identify the better result giving classification algorithm for this dataset.

4.2.3 Experimental Result

Experimental results of our proposed method as elaborated below, can be judged based on the aspects which have been mentioned previously.

Accuracy Percentage

After applying the above mentioned algorithms on the dataset we have found the accuracy result of those algorithms for the dataset and plotted the result in bar chart. Here, the mentioned result of the accuracy are highest after tuning the algorithm parameters value.



```
DecisionTree_accuracy:          86.36%

RandomForest_accuracy:          90.91%

KNN_accuracy:                   72.73%
```

Figure 15: Dataset-2 Model Accuracy

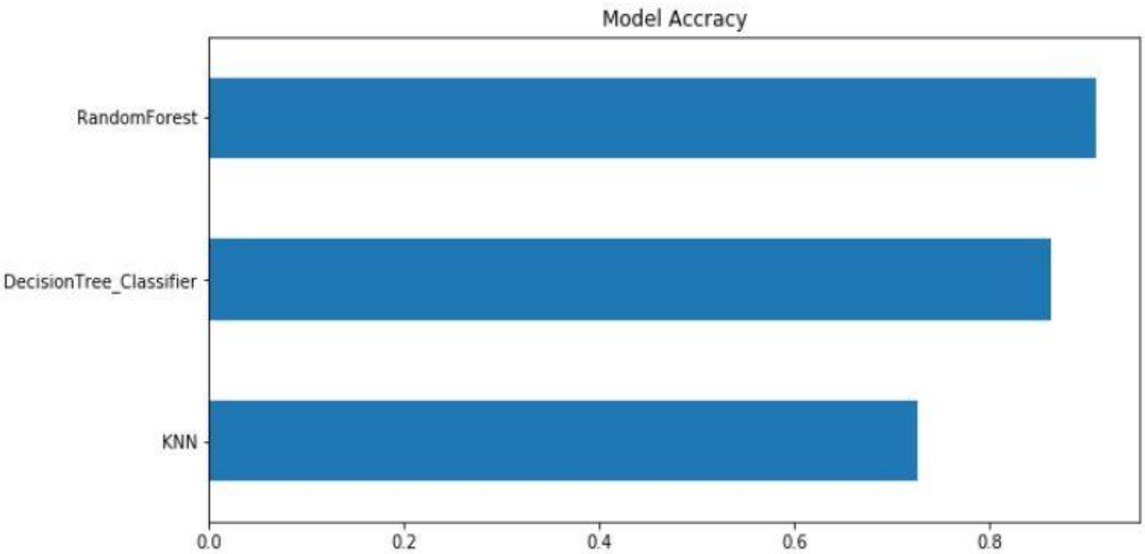
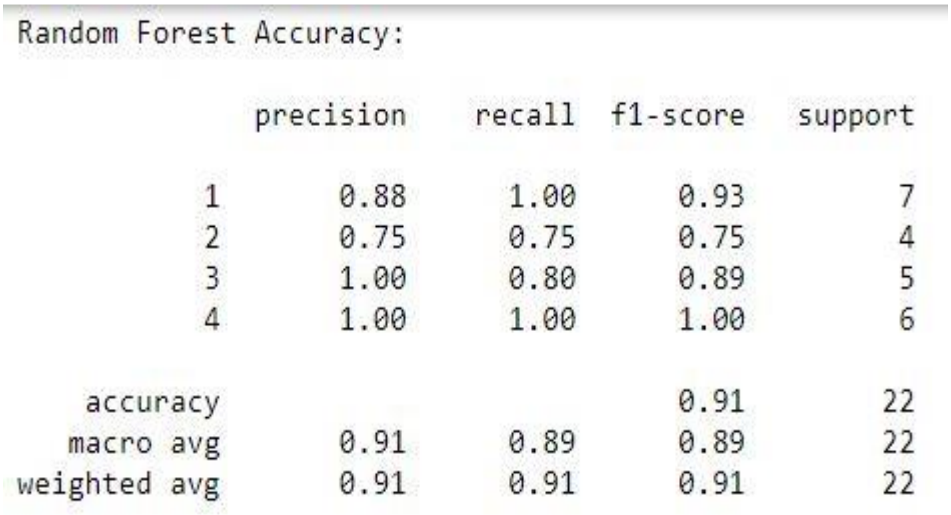


Figure 16: Dataset-2 Model Accuracy Chart

As well as previous sub experiment portion after finding the accuracy percentage of the models we have investigated about precision, recall and f-1 score of the models for getting better insight about the accuracy.



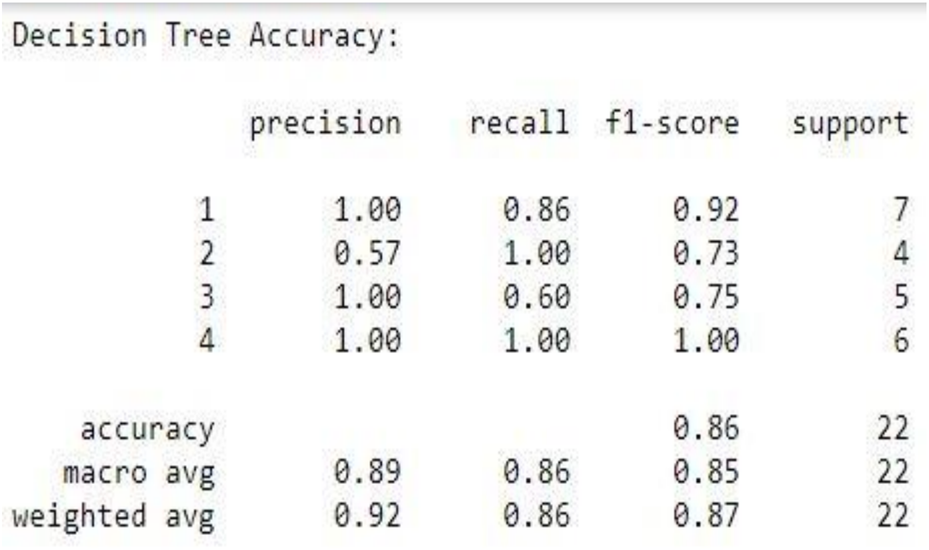
```
Random Forest Accuracy:

              precision    recall  f1-score   support

           1       0.88      1.00      0.93         7
           2       0.75      0.75      0.75         4
           3       1.00      0.80      0.89         5
           4       1.00      1.00      1.00         6

    accuracy                           0.91        22
   macro avg       0.91      0.89      0.89        22
weighted avg       0.91      0.91      0.91        22
```

Figure 17: Dataset-2 Random Forest classification report

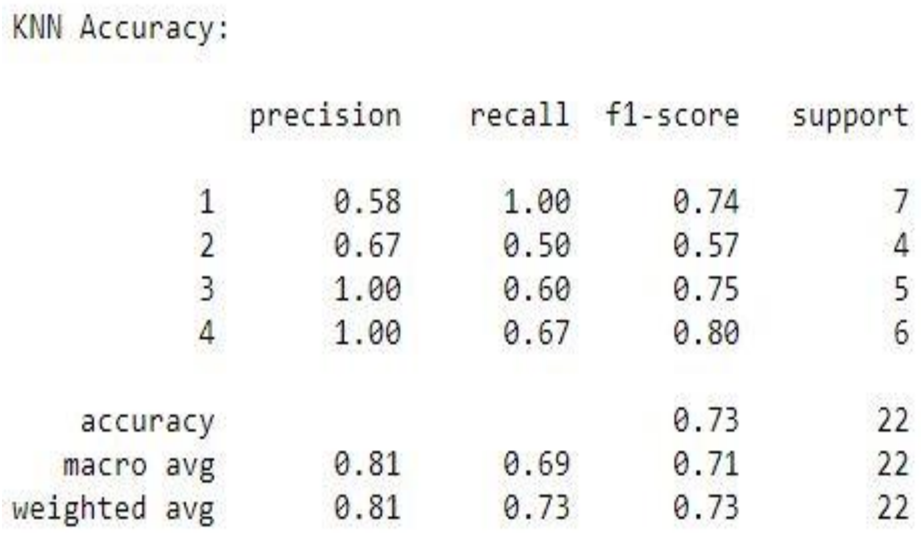


```
Decision Tree Accuracy:
              precision    recall  f1-score   support

           1       1.00      0.86      0.92         7
           2       0.57      1.00      0.73         4
           3       1.00      0.60      0.75         5
           4       1.00      1.00      1.00         6

    accuracy                           0.86        22
   macro avg       0.89      0.86      0.85        22
weighted avg       0.92      0.86      0.87        22
```

Figure 18: Dataset-2 Deciom Tree classification report



```
KNN Accuracy:
              precision    recall  f1-score   support

           1       0.58      1.00      0.74         7
           2       0.67      0.50      0.57         4
           3       1.00      0.60      0.75         5
           4       1.00      0.67      0.80         6

    accuracy                           0.73        22
   macro avg       0.81      0.69      0.71        22
weighted avg       0.81      0.73      0.73        22
```

Figure 19: Dataset-2 KNN classification report

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In our research our main intention was to find:

1. Getting all cluster labels, that ensemble the cluster labels and compare them with our raw data cluster label.
2. Comparing both labels, we can predict the reliability of these Sites sources.
3. Finding from the compare of clusters labels, we'll use the classification system to get better accuracy and gain the prediction correctly.

	Algorithm	Accuracy	Precision	Recall	F-1 Score
Dataset-1	**Decision Tree**	77.27%	70%	77%	73%
	Random Forest	81.82%	69%	82%	75%
	KNN	59.09%	42%	59%	49

Table 1: Dataset-1 Classification report table

	Algorithm	Accuracy	Precision	Recall	F-1 Score
Dataset-2	**Decision Tree**	86.36%	92%	86%	87%
	Random Forest	90.91%	91%	91%	91%
	KNN	72.73%	81%	73%	73%

Table 2: Dataset-2 Classification report table

From dataset-1 and dataset-2 classification report table, we can see that, dataset-2 has better accuracy, precision, recall, f-1 score than dataset-1. That means, performance of partinoed clustering is better than traditional direct clustering technique. It indicates that our proposed approach is far better to detect the credibility of online news portals.

5.1 Future work

In future, we will work on large size of online news portal dataset to find out the credibility. We will apply Neural Network, Deep Neural Network, k-medoids and others supervised and unsupervised learning algorithm and some other optimization method to gain more accuracy.

References

[1] X. a. A. A. a. S. K. A. Wang, "Intelligent web traffic mining and analysis," *Journal of Network and Computer Applications,* vol. 28, pp. 147--165, 2005.

[2] P. a. L. N. Owezarski, "Internet traffic characterization--An analysis of traffic oscillation," in *IEEE International Conference on High Speed Networks and Multimedia Communications*, 2004, pp. 96--107.

[3] S. a. M. U. R. M. a. A. K. M. a. H. N. Dhoju, "Differences in Health News from Reliable and Unreliable Media," in *Companion Proceedings of The 2019 World Wide Web Conference*, 2019, pp. 981--987.

[4] R. K. Kaliyar, "Fake News Detection Using A Deep Neural Network," in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, 2018, pp. 1--7.

[5] Y. a. S. D. a. J. C.-S. Seo, "Fake News Detection Model Using Media Reliability," in *TENCON 2018-2018 IEEE Region 10 Conference*, 2018, pp. 1834--1838.

[6] L. Vaughan, "An alternative data source for web hyperlink analysis:“Sites Linking In” at Alexa Internet," *Collnet journal of scientometrics and information management,* vol. 6, pp. 31--42, 2012.

[7] L. a. I. P. Björneborn, "Toward a basic framework for webometrics," *Journal of the American society for information science and technology,* vol. 55, pp. 1216--1227, 2004.

[8] T. C. a. I. P. Almind, "Informetric analyses on the world wide web: methodological approaches to ‘webometrics’," *Journal of documentation,* vol. 53, pp. 404--426, 1997.

[9] K. a. R. M. S. Naheem, "Webometric analysis of telugu news paper websites: an evaluative study using alexa internet," *International Journal of Digital Library Services,* vol. 7, pp. 26--32, 2017.

[10] K. a. R. M. S. Naheem, "Webometric analysis of telugu news paper websites: an evaluative study using alexa internet," *International Journal of Digital Library Services,* vol. 7, pp. 26--32, 2017.

[11] Y. a. C. N. J. a. R. V. L. Chen, "Misleading online content: Recognizing clickbait as false news," in *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, 2015, pp. 15--19.

[12] S. a. A. V. Shelke, "Source detection of rumor in social network--a review," *Online Social Networks and Media,* vol. 9, pp. 30--42, 2019.

[13] M. Hanief Bhat, "Evaluating Indian newspaper web sites using Alexa Internet," *Library Review,* vol. 62, pp. 398--406, 2013.