## Tribhuvan University

## Institute of Science and Technology

# Comparative Study of Clustering Algorithms for Nepali News

## Dissertation

Submitted to

Central Department of Computer Science & Information Technology

Kirtipur, Kathmandu, Nepal

In partial fulfillment of the requirements

for the Masters Degree in Computer Science & Information Technology

By

**Yub Raj Dahal**

**TU Regd. No.:** 5-2-33-622-2006

Supervisor

**Asst. Prof. Nawaraj Paudel**

March, 2019

# Tribhuvan University

## Institute of Science and Technology

### Central Department of Computer Science & Information Technology

## Students Declaration

I hereby declare that I am the only author of this work and that no sources other than the listed here have been used in this work.

… … … … … … … …

**Yub Raj Dahal**

**Date: March, 2019**

# Tribhuvan University

## Institute of Science and Technology

### Central Department of Computer Science & Information Technology

## Supervisor's Recommendation

I hereby recommend that this dissertation prepared under my supervision by **Mr. Yub Raj Dahal** entitled **Comparative Study of Clustering Algorithms for Nepali News** in partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Information Technology be processed for the evaluation.

… … … … … … … … …

**Asst. Prof. Nawaraj Paudel**
Central Department of Computer Science & IT,
Institute of Science and Technology,
Kirtipur, Kathmandu, Nepal

**Date: March, 2019**

# Tribhuvan University

## Institute of Science and Technology

### Central Department of Computer Science & Information Technology

# LETTER OF APPROVAL

We certify that we have read this dissertation and in our opinion it is satisfactory in the scope and quality as a dissertation in the partial fulfillment for the requirement of Masters Degree in Computer Science and Information Technology.

**Evaluation Committee**

.... .... .... .... .... .... .... ....

**Asst. Prof. Nawaraj Paudel**

Central Department of Computer Science & IT

Tribhuvan University, Kathmandu, Nepal

**(Head of Dept.)**

.... .... .... .... .... .... .... .... ... ... ...

**Asst. Prof. Nawaraj Paudel**

Central Department of Computer Science & IT

Tribhuvan University, Kathmandu, Nepal

**(Supervisor)**

.... .... .... .... .... .... .... ....

**—- —- —- —— —- —-**

**(External Examinar)**

.... .... .... .... .... .... .... .... ...

**—- —- —- —- —- —-**

**(Internal Examinar)**

# ABSTRACT

Clustering is an important technique to separate data categories based on their feature similarity. Clustering belong to unsupervised type of machine learning algorithms. Among many clustering algorithms, three representative algorithms namely K-means, X-means and Expectation Maximization are experimented for the Nepali news clustering problem in this research work. News clustering is the task of categorizing news into groups that share similar interests. Clustering algorithms are evaluated for optimal performances based on cluster evaluation metrics and execution time. Evaluation metrics used are Dunn index, DB index and CH index. Execution time includes clustering time and training time. TF-IDF is used as a news embedding representation. Algorithms are also evaluated with reduced feature dimensions by applying PCA.

To select the winner algorithm and setting the values of DB index, training time and clustering time must be lower and value of CH index and Dunn index must be higher. So, based upon the evaluation results, we conclude the winning algorithm and strategies in some states as follows. When feature dimension is high ($>= 10000$) K-Means perform better then others. When applied PCA to reduce feature space, EM algorithm better performs than others. With reduced feature space, K-Means still performs better then X-Means clustering algorithm.

**Keywords:**

*News Clustering, Natural Language Processing, Nepali language, K-Means, X-Means, EM, PCA*

# ACKNOWLEDGMENTS

I am greatly indebted to my thesis supervisor, Asst. Prof. Nawaraj Paudel, respected Head of Department of Central Department of Computer Science and Information Technology for providing his guidance and support to carry out my dissertation work to the success. At the end of my thesis I would like to thank all those people who made this thesis possible and an unforgettable experience for me. I am very much grateful and thankful to all the respected teachers for providing me such a broad knowledge and inspirations. Special thanks to my family for their endless motivation, constant mental support and love which have been influential in whatever I have achieved so far. All my class fellows are worthy of my gratefulness for their direct or indirect support in completion of my dissertation. I have done my best to complete this research work. Suggestions from the readers are always welcomed, which will improve this work.

Finally, my thanks goes to the authors whose works I have consulted and taken their algorithm to reference.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ALGORITHMS

# LIST OF ABBREVIATIONS

**ML** (Machine Learning)

**KDD** Knowledge Discovery in Database

**OLAP** Online Analytical Processing

**OLTP** Online Transaction Processing

**WWW** World Wide Web

# Chapter 1

# INTRODUCTION

## 1.1 Introduction

In this era of technology, we deal with information collected from various sources on the Internet.The amount of data that is generated on daily basis is increasing tremendously.With the wide use of internet, a large amount of textual documents are present over internet. Available data is in the form of enterprise information systems, digital documents and in personal files. With the increasing size of data proper handling and analysis of data is very crucial. Text mining is being developed to handle the increasing volumes of the text data.

Text mining, also referred to as text data mining, is the process of deriving high-quality information from text. Text clustering, text classification and text categorization are different functionalities of text mining. Clustering is useful in several exploratory pattern analysis, grouping, decision making and machine learning situations. However,in many problems, there is little prior information available about the data and the decision maker must make as few assumptions about the data as possible. Clustering methodology is particularly appropriate for the exploration of interrelationships among the data points to make an assessment of their structure.

On each successive day, the rate of growth in new journals and publications is accelerating. The volume of these documents has made automatic organization and classification an essential element for the advancement of basic and applied research. Most of the recent work on automatic document classification has involved supervised learning techniques such as classification trees, naive Bayes, Support vector machines, neural nets and ensemble methods. Although many existing approaches to document classification can quickly identify the overall area of a document, few of them can rapidly organize documents into the correct sub-field or areas of specialization.

Document clustering or text clustering is the automatic organization of documents into clusters so that the document within a cluster have high similarity in comparison to documents in other clusters.It is divided into two major subcategories, hard clustering and soft clustering. Soft clustering also known as overlapping clustering is again divided into partitioning, hierarchical and frequent itemset-based clustering. Hard clustering compute the hard assignment of a document to a cluster i.e. each document is assigned to exactly one cluster, giving a set of disjoint clusters. Soft clustering compute the soft assignment i.e. each document is allowed to appear in multiple clusters thus, generates a set of overlapping clusters.Partitioning clustering allocate documents into a fixed number of non-empty clusters. Similarly, hierarchical document clustering is to build dendrogram, a hierarchical tree of clusters, whose leaf node represents the subset of a document collection [1]. Hierarchical clustering gives better quality clustering, but is limited because of its quadratic time complexity. Whereas, partitioning methods like K-means and its variants have a linear time complexity, making it more suitable for clustering large datasets but are thought to produce inferior clusters [2].

Clustering is an important means of data mining algorithm that separate data of similar nature. Unlike the classification algorithm, clustering belongs to the unsupervised types of algorithms. Two representatives of the clustering algorithms are the K-means algorithm and the expectation maximization algorithm. EM and K-means are similar in the sense that they allow model refining of an iterative process to find the best congestion. However, the K-means algorithm differs in the method used for calculating Euclidean distance while calculating the distance between each of two data items; and EM uses statistical methods. The EM algorithm is often used to provide the functions more effectively.

The use of communication technologies and and the information content has increased extensively. With the increment in the use of electronic data and the information is stored in electronic format in the form of text documents such as news articles,books,digital library and so on.News articles have been a common source to gain and enhance knowledge, which could be acquired from blogs, online newspapers or news portals. The sources of information might be different but the knowledge they give, however, is of same kind.

People have different interests and expect to retrieve content and its information in various known national or regional language. Growing use of these languages on the Internet has triggered multilingual research which has led to researchers delicately working in this field, consequently exploring various alternatives of it.

With the rapid advancement in technology, we are able to accumulate huge amount of data of numerous kinds. News Clustering also known as document clustering (subset of data clustering) is a technique of data mining which includes concepts from the fields of information retrieval, natural language processing and machine learning. News clustering organizes news into different clusters where news of each cluster share common properties according to defined similarity measures. The fast and high quality clustering algorithm (the algorithm with better semantic representation of the knowledge base) plays a vital role in helping users to effectively navigate, summarize and organize the information.

For a person who frequently reads the news from at least two sources it would be convenient if all those sources news could be read at a single location. It would be more superior in the case with similar articles that only one of them are presented to the users. Furthermore, it would also be desirable if the reader could access all similar news easily. This would be beneficial for source criticism and finding further information on same topic.

News clustering is the problem of grouping news based on their similarity. Similar news appear in the same cluster while while different documents appear in different clusters. Choosing the right function to determine similarity between news is not obvious. Admissible information retrieval and information clustering is an important task in this span of time, as there are immense amount of information in the web. Without vagueness, Nepali news information are used and updated by different users. Selection of words to give information may be different but its content and information is same. Thus, Nepali text clustering is pivotal.

News articles and clustering are widely used over internet for various languages however there is no automated service that aggregates and clusters the Nepali news from various Nepali news agencies. Our system especially works on clustering admissible Nepali news collected from various sources on Internet.

## 1.2 Motivation

Getting closer to the foremost steps of the thesis, experimental study utilizing several algorithms proposed in literature was performed in order to understand if the existed algorithms can be adopted to the peculiarities of the Nepali news clustering. K-Means clustering, X Means clustering and EM clustering are three clustering techniques that are commonly used for document clustering or news clustering. To experiment the baseline approached for the Nepali news clustering problem we choose these three clustering algorithms as main algorithms.

## 1.3 Problem Definition

A person who is reading particular news in one news portal might also be interested in reading similar news coverage on other news portal to find more about the topic. The problem here is finding the news portal that covers the similar news. The usual approach is to visit each likely news portal and then manually look for the similar news in them to find whether the news the person is looking for is present or not. This is problematic, time-consuming and a tedious task. This might even reduce the users interest in reading that particular news as well as his enthusiasm to acquire more information on that topic. On the other hand, people are so much into smart technologies these days that they always look for the technology that can satisfy their interest without having them to put in much effort and time.

The solution proposed here is based on the idea of text mining and clustering. The basic idea is to create clusters of similar news headlines and build news content warehouse.

The statement of this thesis is to experiment baseline clustering algorithms for efficiency and accuracy of the Nepali news clustering problem. We also apply variations on text feature embedding generations to get the effectiveness of the feature selection approaches.

## 1.4 Objectives

The main objective of this dissertation work are as follows:

- To compare the performance and efficiency of the K-Means, X-Means and EM clustering algorithms for Nepali news clustering problem.

## 1.5 Outline of the Report

The thesis document is organized as follows:

Chapter 1 describes the thesis statement and objectives.

Chapter 2 describes the background and overview of the data mining approaches.

Chapter 3 describes the state of the art of the clustering algorithms and researches in the news clustering.

Chapter 4 describes about the methodology and algorithm used.

Chapter 5 provides information of the datasets used for the research.

Chapter 6 contains the experimental results done on datasets using the described methodology algorithms.

Chapter 7 contains the summary and future scope of the research work.

# Chapter 2

# Background and Overview

## 2.1 Data mining

Over the past two decades there has been a huge increase in the amount of data being stored in database as well as the number of database applications in business and the scientific domain. The huge amounts of stored data contains knowledge about the number of aspects of their business waiting to be harnessed and used for more effective business decision support. Database Management systems used to manage these data sets at present only allow the user to access information explicitly present in the databases. The data stored in database is only a small part of ”iceberg of information” available from it. Contained implicitly within this data is knowledge about a number of aspects of their business waiting to be harnessed and used for more effective business decision support. This extraction of knowledge from large data sets is called Data Mining or Knowledge Discovery in databases and is defined as the non-trivial extraction of implicit, previously unknown and potentially useful information from data [3].

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses [4]. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining can be considered to be an inter-disciplinary field involving concepts from machine learning, database technology, statistics, mathematics, clustering and visualization among others.

Data mining is about learning from existing real-world data rather than data generated particularly for the learning tasks. In data mining the data sets are large therefore efficiency and scalability of algorithms is important. Almost in parallel with the developments in the database field, machine learning research was maturing with the development of a number of sophisticated techniques based on different models of human learning. Learning by example, cased-based reasoning, learning by observation and neural networks are some of the most popular learning techniques that were being used to create the ultimate thinking machine.

### 2.1.1 Functionalities of data mining

#### 2.1.1.1 Outlier Analysis

Outlier analysis is an object in database which is significantly different from the existing data. ”An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism ”. The outlier can be diagnosed with the help of statistical tests that assume probability model for the data.

#### 2.1.1.2 Evolution Analysis

Evolution analysis is the mechanism of extracting pattern from data changes over time.

#### 2.1.1.3 Association Analysis

Association is a data mining function that discovers the probability of the co-occurrence of items in a collection. The relationships between co-occurring items are expressed as association rules.

#### 2.1.1.4 Clustering

Clustering is the process of partitioning a set of object or data in a same group called a cluster. These objects are more similar to each other than to those in other groups. Clustering is used in many fields including machine learning, pattern recognition, bioinformatics, image analysis and information retrieval.

#### 2.1.1.5 Classification

Classification is used to build models from data with predefined classes as the model is used to classify new instance whose classification is not known. The instances used to create the model are known as training data. A decision tree or set of classification rules is based on such type of mechanism of classification which can be retrieved for identification of future data.

## 2.2 Clustering

Clustering is the grouping of a particular shape of objects based on their characteristics, aggregating them according to their similarities. Regarding to data mining, this methodology partitions the data implementing a specific join algorithm, most suitable for the desired information analysis. The clustering analysis allows an object not to be part of a cluster, or strictly belong to it, known as hard clustering . Similarly, soft partitioning states that every object belongs to a cluster in a determined degree. More specific divisions can be possible to create like objects belonging to multiple clusters, to force an object to participate in only one cluster of even construct hierarchical trees on group relationships.

There are several different ways to implement this partitioning, based on distinct models. Distinct algorithms are applied to each model, differentiating it's properties and results. These models are distinguished by their organization and type of relationship between them.

### 2.2.1 Centroid-based model

In this type of grouping method, every cluster is referenced by a vector of values. Each object is part of the cluster whose values difference is minimal, comparing to other clusters. The number of clusters should be pre-defined, and this is the biggest problem of this kind of algorithms. This methodology is the most close to the classification subject and are vastly used for optimization problems.

### 2.2.2 Distributed-based model

Related to pre-defined statistical models,the distributed methodology combines objects whose values belongs to the same distribution. Because of its random nature of value generation, this process needs a well defined and complex model to interact in a better way with real data. However these processes can achieve a optimal solution and calculate correlations and dependencies.

### 2.2.3 Connectivity-based model

On this type of algorithm, every object is related to its neighbors, depending the degree of that relationship on the distance between them. Based on this assumption, clusters are created with near by objects, and can be described as a maximum distance limit. With this relationship between members, these clusters have hierarchical representations. The distance function varies on the focus of the analysis.

### 2.2.4 Density-based model

These algorithms create clusters according to the high density of members of a data set, in a determined location. It aggregates some distance notation to a density standard level to group members in clusters. These kind of processes may have less performance on detecting the limit areas of the group.

### 2.2.5 Requirements of clustering in data mining

- Scalability : We need highly scalable clustering algorithms to deal with large databases.
- Ability to deal with different kinds of attributes: Algorithms should be capable to be applied on any kind of data such as interval-based data, categorical and binary data.
- Discovery of clusters with attribute shape: The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical clusters of small sizes.
- High Dimensionality: The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
- Ability to deal with noisy data: Database contain noisy, missing or errorneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
- Interpretablility : The clustering results should be interpretable, comprehensible, and usable.

### 2.2.6 Terminologies of clustering

- Cluster : A collection of one or more master and data nodes.
- Master Node: Coordinator of the cluster. Manages the distribution of shards and keeps track of all nodes in the cluster. There can be more than one master node. If a master

node fails, then a new node that is marked as a master node is automatically elected as a new master node. The cluster cannot operate without at least one master node.

- Data Node: Workhorse of the cluster. Stores data and processes incoming search and index requests. A node can act both as a data node and master node.
- Shard: Each data node stores data in a shard.
- Replica Shard: Each Shard can have any number of replicas. Replicas are used to ensure high availability in the case that a node is no longer available.

# Chapter 3

# Literature Review

Clustering is the task of dividing the population or data points into a number of groups such that data point in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters [5].

[6] defines clustering as a concise model of the data which can be interpreted in the sense of either a summary or a generative method. Probabilistic methods, distance-based methods, density-based methods, grid based methods, factorization techniques and spectral techniques are the classes of methods. The diversity of different data types significantly adds to the richness of the clustering problems.Many variations and enhancements of clustering such as visual methods, ensemble methods, multiview methods or supervised methods can be used to improve the quality of the insights obtained from the clustering process.

[7] has proposed a novel document partitioning method based on the non-negative factorization of the term-document matrix of the given document corpus. The proposed document clustering method surpasses singular vector decomposition(SVD) and the eigen decomposition clustering methods not only in the easy and reliable derivation of document clustering results, but also in document clustering accuracies. [8] explores a simple and efficient baseline for text classification.Experiments shows that fast text classifier fastText is often on par with deep learning classifiers in terms of accuracy, and many orders of magnitude faster for training and evaluation.

Document clustering is automatic organization of documents into clusters so that documents within a cluster have high similarity in comparison to documents in other clusters. It has been studied intensively because of its wide applicability in various areas such as web mining, search engines, and information retrieval. It is measuring similarities between documents and grouping similar documents together. It provides efficient representation and visualization of the documents. Clustering is useful in several exploratory pattern-analysis, grouping, decision-making and machine-learning situations, including data mining, document retrieval, image segmentation and pattern classification.

Multi view point based clustering methods with similarity measure by using incremental algorithm approach for clustering high dimensional data is explained by [9]. Existing clustering algorithms are implemented based on partitioning, hierarchical, density based and grid based. All clustering methods have to assume some cluster relationship among the data objects that they are applied on. Similarity between a pair of objects can be defined either explicitly or implicitly.

An overview of pattern clustering methods from a statistical pattern recognition perspective, with a goal of providing useful advice and references to fundamental concepts accessible to the broad community of clustering practitioners is given by [10]. Clustering is a difficult problem combinatorially, and differences in assumption and contexts in different communities has made the transfer of useful generic concepts and methodologies to occur slowly.

[11] emphasizes challenges of the clustering methods in dealing with problems of high dimensionality, scalability, accuracy and meaningful cluster labels. A brief summary over methods studied and current state of documents clustering research, document representation model and its challenges, dimensionality reduction mechanisms, issues in document clustering and cluster quality evaluation criteria are discussed. In [12], the medical text classification problem is addressed using the convolutional neural networks. The convolutional neural network with medical data set consisting of several classes of health information is trained and tested with the accuracy of about 15% more than the existing approach in this field.

[13] Presents the results of an experimental study of some common document clustering techniques. Two main approaches of document clustering, agglomerative hierarchical clustering and K-means are compared. Results indicate that the bisecting K-means technique is better than the standard K-means approach and as good or better than the hierarchical approaches. In addition, the run time of bisecting K-means is very attractive when compared to that of agglomerative hierarchical clustering techniques.

[14] had spent plenty of time in automatic document clustering using topic analysis. Topic segmentation is applied to detect topics within documents and using term relationships attempt to build hierarchies which represents a ”real world” topic hierarchy.Documents are assigned to each of these topics using a standard clustering techniques. Two methods for document clustering systems has been proposed. Foremost, an adaptation of tree measure algorithms to document hierarchies which requires a predefined tree which has been agreed upon as a suitable benchmark. The later is independent of any benchmark trees and presents the evaluator with

the number of measures which allow to assess the properties of the tree.

Comparison of various clustering algorithms are done in [15]. Every algorithm has their own importance and we use them on the behavior of the data, but on the basis of performed research k-means clustering algorithm is simplest algorithm as compared to other algorithm. In [16] analysis of algorithm and comparing the various clustering algorithm by using WEKA tool to find out which algorithm will be more comfortable for the users for performing clustering algorithm. This presents the application's of data mining WEKA tool and provides huge clusters of data set which can be used for search engine optimization.

A recurrent convolutional neural network for text classification without human designed features is discussed in [17]. A recurrent structure to capture contextual information as far as possible when learning word representations, which may introduce considerably less noise compared to traditional window-based neural networks. A max-pooling layer is employed that automatically judges with words play key roles in text classification to capture the key components in texts. A different approach from current document classification methods that view the problem as multi-class classification is introduced in [18]. Approach known as Hierarchical deep learning for text classification employs stacks of deep learning architectures to provide specialized at each level of the document hierarchy.

The comparison of algorithms K-Means and Expectation-Maximization in clustering is discussed in [19]. The effectiveness of Expectation-Maximization clustering algorithm is measured through a number of internal and external validity metrics, stability, runtime and scalability tests. An algorithm that efficiently searches the space of cluster locations and number of clusters to optimize the Bayesian information criterion and Alkaike information criterion measure is introduced by [20]. Experiments shows that proposed technique reveals the true number of classes in the underlying distribution and is much faster and repeatedly using accelerated K-Means for different values of K.

A general approach to iterative computation of maximum-likelihood estimates when the observation can be viewed as incomplete data is presented in [21]. Each iteration of the algorithm consists of an expectation step followed by a maximization step known as EM algorithm. When the underlying, complete data come from an exponential family whose maximum-likelihood estimates are easily computed, then each maximization step of an EM algorithm is likewise easily computed. A general classification EM algorithm is defined in [22]. Numerical experiments, reported for the variance criterion, show that both stochastic algorithms perform well compared with the standard K-means algorithm which is particular version of the classification EM algorithm.

Machine generated decision rules appear comparable to human performance, while using the identical rule-based representation [23]. Human engineered systems, using the identical representation of production rules, can be successful in text classification. Rule-based systems for text classification can be automatically generated from samples with very comparable performance measures. The selection of the best classification algorithm for a given dataset is very widespread problem, occurring each time one has to choose a classifier to solve a real-world problem [24]. One of the most crucial, is the choice of an appropriate measure in order to properly assess the classification performance and rank the algorithms.

[25] explores the use of Support vector machines for learning text classifiers from example.The hierarchical topic structure to decompose the classification task into a set of simpler problems, one at each node in the classification is discussed in [26]. The finding of the experiments shows the accuracy of flat classification decreases as the number of classes and documents increases.

Optimal part-of-speech tagging have great importance in various field of natural language processing such as machine translation, information extraction, word sense disambiguation, speech recognition and others. Due to the speech nature of the Nepali language, Tagset used and size of the corpus getting accurate part of speech tagger is one of the challenging task [27]. POS tagging of Nepali is a necessary component for most NLP applications in Nepali, which analyses the construction of the language, behavior of the language and can be used to develop automated tools for language processing [28].

The lack of a standard Nepali corpus prompted for the creation of a manual data set by crawling various Nepali news sites [29]. Nepali document classification is severely limited by the complexity of the language morphology. Document classification with word2vec employs neural network and simplifies the process of automatically categorizing Nepali documents while increasing the precision and recall over previously implemented techniques such as TF-IDF.

The concept of traditional rule based system and corpus based approach is tested for Nepali language which is based in devanagarik script [30]. The approach has given better result in comparison to traditional rule based system particularly for Nepali language. The strength and weakness of support vector machine based named entity recognition for using Nepali text is discussed in [31]. Development of stemming tool, part of speech tagging and Named entity recognition detection tool using semi hybrid approach and some rule based approaches and its accuracies is defined in [32].

The implementation of the Naive Bayes and SVM-based classification techniques to classify the Nepali SMS as Spam and non-spam and evaluation for accuracy measure of the classification methodologies is done in [33]. [34] explains the evaluation of lexicon-pooled Naive Bayes approach by applying sufficiently large datasets of Nepalese news stories and its accuracies and usefulness of the method for Nepali news classification.

[35] has reviewed various algorithms with a wide range of approaches to solve problem of news clustering. A hierarchical algorithm that incorporates various ideas of researchers have been implemented. It is concluded that fuzzy equivalence algorithm does produce acceptable results when compared to Google News as a reference. The algorithm however requires a huge amount of memory to hold the trained model. This render is not suitable to run on portable devices but very suitable to run on a server farm. Moreover, during the training phase, the algorithm builds and the memory and that makes it hard to process the full training dataset on a single processing node.

In [14], two evaluation methods for document clustering system is introduced. The first is an adaptation of tree major algorithms to document hierarchies requiring a pre-defined tree which has been agreed upon as a suitable benchmark. The second is independent of any benchmark trees and presents the evaluator with a number of measures which allows to assess the properties of the tree. Internal clustering validation and and a detail study of 11 widely used internal clustering validation measures for crisp clustering is elaborated by [36]. Experimentation shows that *SDbw* is the only internal validation measure which performs well in any five aspects, while other measures have certain limitations in different application scenarios.

Automated news classification is the task of categorizing news into some predefined category based on their content with the confidence learned from training news dataset.Evaluation of most widely used machine learning techniques, mainly Naive Bayes, SVM and neural networks for automatic Nepali news classification problem is discussed in [37].

Overview of various document clustering methods, starting from basic traditional methods to fuzzy based, genetic, co-clustering, heuristic oriented etc., and the document clustering procedure with feature selection process, applications, challenges in document clustering, similarity measures and evaluation of document clustering algorithm is explained in the paper [38].

# Chapter 4

# Research Methodology

### 4.0.1 System Overview

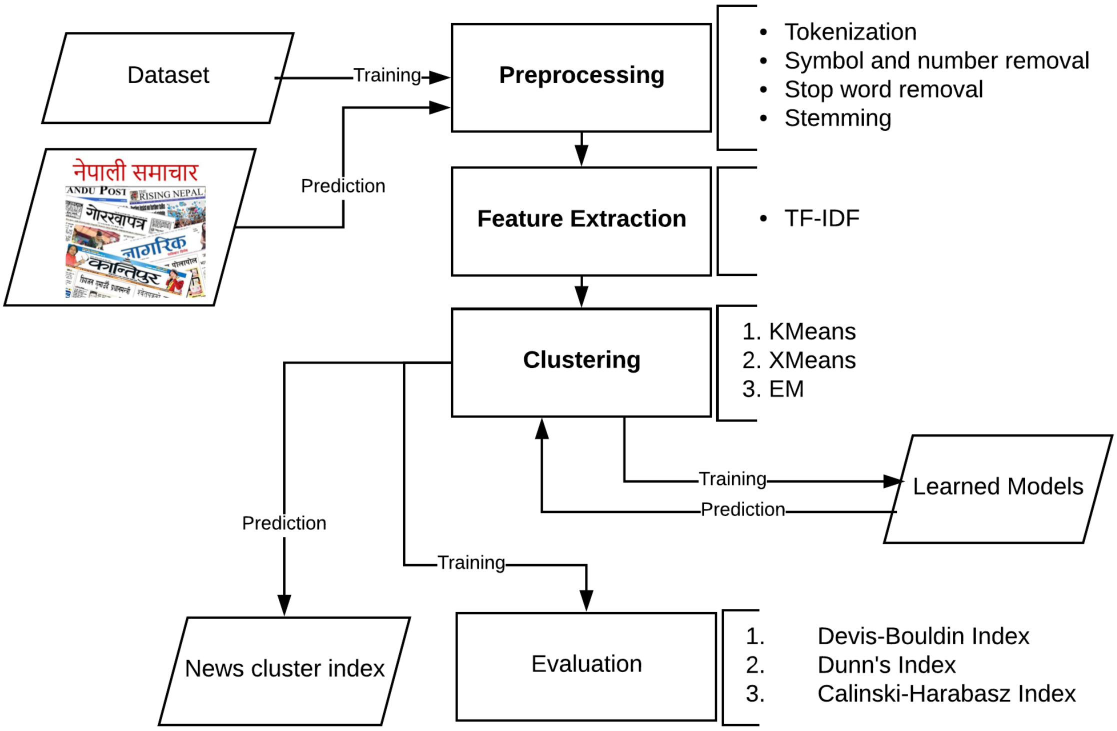


Figure 4.1: High level system diagram

### 4.0.2 TF-IDF Features

TF-IDF (Term Frequency-Inverse Document Frequency) is a text mining technique used to categorize documents. This algorithm is useful when document set is large and needs to be categorized. It is especially nifty because training a model ahead of time is not required and will automatically account for differences in lengths of documents.

TF-IDF computes a weight which represents the importance of a term inside a document. It does this by comparing the frequency of usage inside an individual document as opposed to the entire data set.

$$TF(t) = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document}$$

.

$$IDF(t) = \log_e \frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ t\ in\ it}$$

.

$$TF\_IDF = TF * IDF$$

.

IF-IDF is computed for each term in each document. It can be done either in one term in particular or in terms with the highest TF-IDF in a specific document.

To convert Nepali news document into it's numerical representation using TF-IDF approach, we follow the following steps.

- Create vocabulary by scanning the training dataset.
- Sort the vocabulary based upon term frequency in descending order.
- Remove the stopwords
- Select topK vocabulary as a feature dimension.
- Scan the news document and preprocess it.
- Scan the preprocessed document and populate the feature of corresponding term in the feature row by calculating the TF-IDF value.

### 4.0.3 K-Means Clustering

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters(assume *k* clusters) fixed apriori by minimizing the average squared distance between points in the same cluster. The k-means algorithm [2] is given below.

**Algorithm 4.1** The k-means algorithm

```
Arbitrarily choose an initial $k$ centers $C = c_1, c_2, ..., c_k$.
For each $i \in 1, ..., k$, set the cluster $C_i$ to be the set of points in $X$ that are closer to $c_i$ than they are to $c_j$ for all $j \neq i$.
For each $i \in 1, ..., k$, set $c_i$ to be the center of mass of all points in $C_i : c_i = \frac{1}{C_i} \sum_{x \in C_i} x$.
Repeat Steps 2 and 3 until $C$ no longer changes.
```

### 4.0.4 X-Means Clustering

X-Means clustering algorithm, an extended K-Means which tries to automatically determine the number of clusters based on BIC scores. Starting with only one cluster, the X-Means algorithm goes into action after each run of K-Means, making local decisions about which subset of the current centroids should split themselves in order to better fit the data. The splitting decision is done by computing the Bayesian Information Criterion (BIC) [20].

### 4.0.5 EM Clustering

EM assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters. EM can decide how many clusters to create by cross validation, or you may specify apriori how many clusters to generate.

The cross validation performed to determine the number of clusters is done in the following steps:

**Algorithm 4.2** Cluster center selection algorithm

```
The number of clusters is set to 1.
The training set is split randomly into $n = 10$ folds.
EM is performed $n$ times using the $n$ folds the usual cross-validation way.    ▷ EM Algorithm 4.3
The log-likelihood is averaged over all $n$ results.
If log likelihood has increased the number of clusters is increased by 1 and the program continues at step 2.
```

The number of folds is fixed to $n$, as long as the number of instances in the training set is not smaller $n$. If this is the case the number of folds is set equal to the number of instances.

An expectation maximization (EM) [21] algorithm is an iterative method for finding maximum likelihood or maximum a posteriori estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM [22] iteration alternates between performing an expectation (E) step, which computes the expectation of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

**Algorithm 4.3** EM algorithm

```
1: Expectation step: For each data point $x$, compute the membership probability of $x$ in each cluster $c_1, c_2, ..., c_k$.
2: Maximization step: Update mixture model parameter (probability weight).
3: Stopping criteria: If stopping criteria is not satisfied, goto step 1.
```

### 4.0.6 Principal Component Analysis

Principal Component Analysis(PCA) is one of the important algorithms in the field of Data Science and is by far the most popularly dimensionality reduction method currently used today.The objective of PCA is simple, identify a hyperplane that lies closest to the data points, and project the data onto it.

The main idea of principal component analysis is to reduce the dimensionality of data set consisting of many variables correlated with each other,either heavily or lightly,while retaining the variation present in the dataset, up to the maximum extent.The same is done by transforming the variables to a new set of variables, which are known as principal components and are orthogonal, ordered such that the retention of variation present in the original variables decreases as we move down in the order. So, in this way, the 1st principal component retains maximum variation that was present in the original components. The principal components are the eigenvectors of a covariance matrix, and hence they are orthogonal.

## 4.1 Cluster Evaluation

A good clustering method will produce high quality clusters in which the intra-class(that is, intra-cluster) similarity is high and the inter-class similarity is low [36]. The quality of a clustering result also depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns. However, objective evaluation is problematic: usually done by human / expert inspection.

### 4.1.1 The Davies-Bouldin Index

The Davies-Bouldin (DB) index [39] is calculated as follows. For each cluster $c$, the similarities between $c$ and all other clusters are computed, and the highest value is assigned to $c$ as its cluster similarity. Then the DB index can be obtained by averaging all the cluster similarities. The smaller the index is, the better the clustering result is. By minimizing this index, clusters are the most distinctfrom each other, and therefore achieves the best partition.

$$DB = \frac{1}{NC}\sum_{i}\{max_{j,j\neq i}\{[\frac{1}{n_i}\sum_{x\in C_i} d(x,c_i) + \frac{1}{n_j}\sum_{x\in C_j} d(x,c_j)]/d(c_i,c_j)\} \quad (4.1)$$

Where,

$NC$: number of clusters,

$n_i$: number of points in $C_i$,

$C_i$: the $i^{th}$ cluster,

$d(x,y)$: distance between $x$ and $y$

### 4.1.2 Dunns Index

Dunn’s index [40] uses the minimum pairwise distance between objects in different clusters as the inter-cluster separation and the maximum diameter among all clusters as the intra-cluster compactness. The optimal cluster number is determined by maximizing the value of the index.

$$D = min_i\{min_j(\frac{min_{x\in C_i, y\in C_j} d(x,y)}{max_k\{max_{x,y\in C_k} d(x,y)\}})\} \quad (4.2)$$

Where,

$C_i$: the $i^{th}$ cluster,

$d(x,y)$: distance between $x$ and $y$

### 4.1.3 The Calinski-Harabasz Index

The Calinski-Harabasz index (CH) [41] evaluates the cluster validity based on the average between- and within-cluster sum of squares. Well separated and compact clusters should maximize this ratio.

$$CH = \frac{\sum_i n_i d^2(c_i, c)/(NC-1)}{\sum_i \sum_{c \in C_i} d^2(x, c_i)/(n-NC)} \tag{4.3}$$

Where,

$NC$: number of clusters,

$n_i$: number of points in $C_i$,

$C_i$: the $i^{th}$ cluster,

$d(x,y)$: distance between $x$ and $y$

# Chapter 5

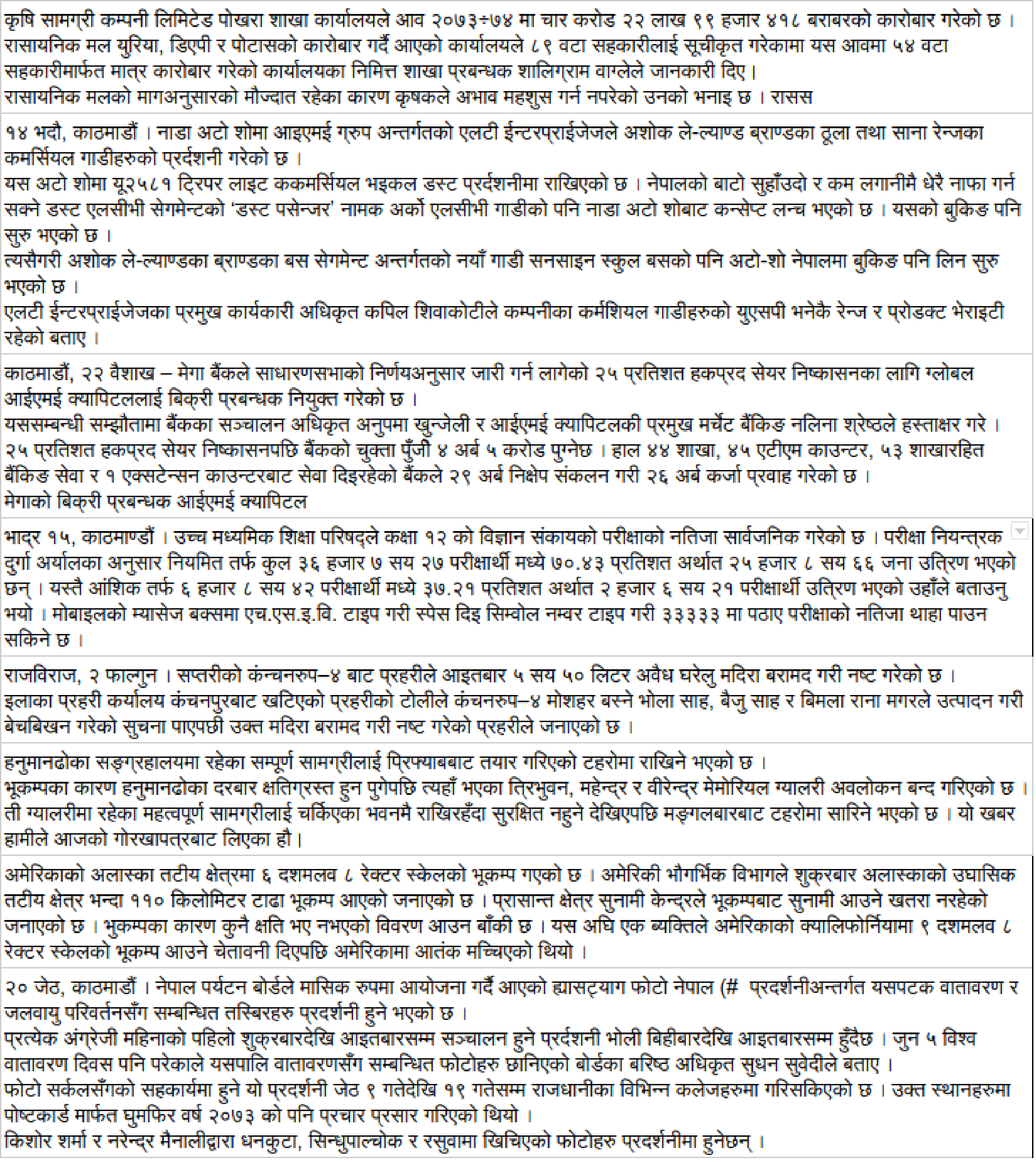
# Datasets

To evaluate the system, Nepali news dataset (Table 5.1) is taken from paper [33]. They have collected it from various online sources , mainly Nepali news portals (e.g., ratopati.com,setopati.com, onlinekhabar.com, and ekantipur.com), using the web crawler. It contains total 4964 news from different domains. For clustering purpose news documents are feed to the system without labels.

Table 5.1: Nepali news dataset

| S.N. | News class | No. of documents |
|---|---|---|
| 1 | Agriculture | 100 |
| 2 | Automobile | 95 |
| 3 | Bank | 417 |
| 4 | Blog | 209 |
| 5 | Business | 142 |
| 6 | Economy | 500 |
| 7 | Education | 85 |
| 8 | Employment | 154 |
| 9 | Entertainment | 500 |
| 10 | Health | 31 |
| 11 | Interview | 229 |
| 12 | Literature | 102 |
| 13 | Migration | 111 |
| 14 | Opinion | 500 |
| 15 | Politics | 500 |
| 16 | Society | 253 |
| 17 | Sport | 500 |
| 18 | Technology | 110 |
| 19 | Tourism | 214 |
| 20 | World | 212 |
| | **Total** | **4,964** |

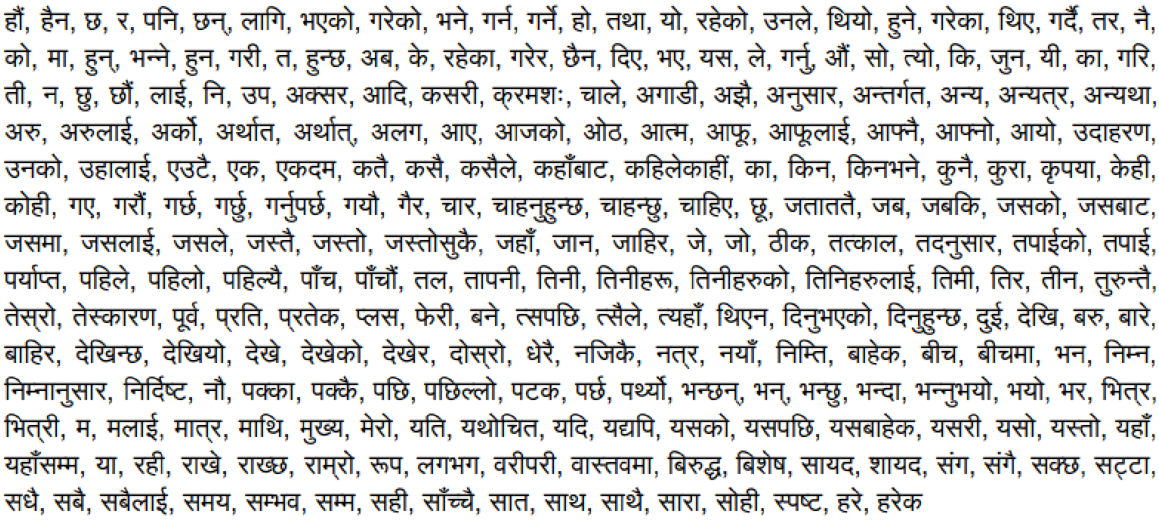
Sample Nepali news are given in the Figure 5.1.



| |
|---|
| कृषि सामग्री कम्पनी लिमिटेड पोखरा शाखा कार्यालयले आव २०७३÷७४ मा चार करोड २२ लाख ९९ हजार ४१८ बराबरको कारोबार गरेको छ ।<br>रासायनिक मल युरिया, डिएपी र पोटासको कारोबार गर्दै आएको कार्यालयले ८९ वटा सहकारीलाई सूचीकृत गरेकामा यस आवमा ५४ वटा सहकारीमार्फत मात्र कारोबार गरेको कार्यालयका निमित्त शाखा प्रबन्धक शालिग्राम वाग्लेले जानकारी दिए ।<br>रासायनिक मलको मागअनुसारको मौज्दात रहेका कारण कृषकले अभाव महशुस गर्न नपरेको उनको भनाइ छ । रासस |
| १४ भदौ, काठमाडौं । नाडा अटो शोमा आइएमई ग्रुप अन्तर्गतको एलटी ईन्टरप्राईजेजले अशोक ले-ल्याण्ड ब्राण्डका ठूला तथा साना रेन्जका कमर्सियल गाडीहरुको प्रर्दशनी गरेको छ ।<br>यस अटो शोमा यू२५८१ ट्रिपर लाइट ककमर्सियल भइकल डस्ट प्रर्दशनीमा राखिएको छ । नेपालको बाटो सुहाँउदो र कम लगानीमै धेरै नाफा गर्न सक्ने डस्ट एलसीभी सेगमेन्टको ‘डस्ट पसेन्जर’ नामक अर्को एलसीभी गाडीको पनि नाडा अटो शोबाट कन्सेप्ट लन्च भएको छ । यसको बुकिङ पनि सुरु भएको छ ।<br>त्यसैगरी अशोक ले-ल्याण्डका ब्राण्डका बस सेगमेन्ट अन्तर्गतको नयाँ गाडी सनसाइन स्कुल बसको पनि अटो-शो नेपालमा बुकिङ पनि लिन सुरु भएको छ ।<br>एलटी ईन्टरप्राईजेजका प्रमुख कार्यकारी अधिकृत कपिल शिवाकोटीले कम्पनीका कर्मशियल गाडीहरुको युएसपी भनेकै रेन्ज र प्रोडक्ट भेराइटी रहेको बताए । |
| काठमाडौं, २२ वैशाख – मेगा बैंकले साधारणसभाको निर्णयअनुसार जारी गर्न लागेको २५ प्रतिशत हकप्रद सेयर निष्कासनका लागि ग्लोबल आईएमई क्यापिटललाई बिक्री प्रबन्धक नियुक्त गरेको छ ।<br>यससम्बन्धी सम्झौतामा बैंकका सञ्चालन अधिकृत अनुपमा खुन्जेली र आईएमई क्यापिटलकी प्रमुख मर्चेट बैंकिङ नलिना श्रेष्ठले हस्ताक्षर गरे । २५ प्रतिशत हकप्रद सेयर निष्कासनपछि बैंकको चुक्ता पुँजी ४ अर्ब ५ करोड पुग्नेछ । हाल ४४ शाखा, ४५ एटीएम काउन्टर, ५३ शाखारहित बैंकिङ सेवा र १ एक्सटेन्सन काउन्टरबाट सेवा दिइरहेको बैंकले २९ अर्ब निक्षेप संकलन गरी २६ अर्ब कर्जा प्रवाह गरेको छ ।<br>मेगाको बिक्री प्रबन्धक आईएमई क्यापिटल |
| भाद्र १५, काठमाण्डौं । उच्च मध्यमिक शिक्षा परिषद्ले कक्षा १२ को विज्ञान संकायको परीक्षाको नतिजा सार्वजनिक गरेको छ । परीक्षा नियन्त्रक दुर्गा अर्यालका अनुसार नियमित तर्फ कुल ३६ हजार ७ सय २७ परीक्षार्थी मध्ये ७०.४३ प्रतिशत अर्थात २५ हजार ८ सय ६६ जना उतिर्ण भएको छन् । यस्तै आंशिक तर्फ ६ हजार ८ सय ४२ परीक्षार्थी मध्ये ३७.२१ प्रतिशत अर्थात २ हजार ६ सय २१ परीक्षार्थी उतिर्ण भएको उहाँले बताउनु भयो । मोबाइलको म्यासेज बक्समा एच.एस.इ.वि. टाइप गरी स्पेस दिइ सिम्वोल नम्वर टाइप गरी ३३३३३ मा पठाए परीक्षाको नतिजा थाहा पाउन सकिने छ । |
| राजविराज, २ फाल्गुन । सप्तरीको कंन्चनरुप–४ बाट प्रहरीले आइतबार ५ सय ५० लिटर अवैध घरेलु मदिरा बरामद गरी नष्ट गरेको छ ।<br>इलाका प्रहरी कर्यालय कंचनपुरबाट खटिएको प्रहरीको टोलीले कंचनरुप–४ मोशहर बस्ने भोला साह, बैजु साह र बिमला राना मगरले उत्पादन गरी बेचबिखन गरेको सुचना पाएपछी उक्त मदिरा बरामद गरी नष्ट गरेको प्रहरीले जनाएको छ । |
| हनुमानढोका सङ्ग्रहालयमा रहेका सम्पूर्ण सामग्रीलाई पिरफ्याबबाट तयार गरिएको टहरोमा राखिने भएको छ ।<br>भूकम्पका कारण हनुमानढोका दरबार क्षतिग्रस्त हुन पुगेपछि त्यहाँ भएका त्रिभुवन, महेन्द्र र वीरेन्द्र मेमोरियल ग्यालरी अवलोकन बन्द गरिएको छ । ती ग्यालरीमा रहेका महत्वपूर्ण सामग्रीलाई चर्किएका भवनमै राखिरहँदा सुरक्षित नहुने देखिएपछि मङ्गलबारबाट टहरोमा सारिने भएको छ । यो खबर हामीले आजको गोरखापत्रबाट लिएका हौ । |
| अमेरिकाको अलास्का तटीय क्षेत्रमा ६ दशमलव ८ रेक्टर स्केलको भूकम्प गएको छ । अमेरिकी भौगर्भिक विभागले शुक्रबार अलास्काको उघासिक तटीय क्षेत्र भन्दा ११० किलोमिटर टाढा भूकम्प आएको जनाएको छ । प्रासान्त क्षेत्र सुनामी केन्द्रले भूकम्पबाट सुनामी आउने खतरा नरहेको जनाएको छ । भुकम्पका कारण कुनै क्षति भए नभएको विवरण आउन बाँकी छ । यस अघि एक ब्यक्तिले अमेरिकाको क्यालिफोर्नियामा ९ दशमलव ८ रेक्टर स्केलको भूकम्प आउने चेतावनी दिएपछि अमेरिकामा आतंक मच्चिएको थियो । |
| २० जेठ, काठमाडौं । नेपाल पर्यटन बोर्डले मासिक रुपमा आयोजना गर्दै आएको ह्यासट्याग फोटो नेपाल (#  प्रदर्शनीअन्तर्गत यसपटक वातावरण र जलवायु परिवर्तनसँग सम्बन्धित तस्बिरहरु प्रदर्शनी हुने भएको छ ।<br>प्रत्येक अंग्रेजी महिनाको पहिलो शुक्रबारदेखि आइतबारसम्म सञ्चालन हुने प्रर्दशनी भोली बिहीबारदेखि आइतबारसम्म हुँदैछ । जुन ५ विश्व वातावरण दिवस पनि परेकाले यसपालि वातावरणसँग सम्बन्धित फोटोहरु छानिएको बोर्डका बरिष्ठ अधिकृत सुधन सुवेदीले बताए ।<br>फोटो सर्कलसँगको सहकार्यमा हुने यो प्रदर्शनी जेठ ९ गतेदेखि ११ गतेसम्म राजधानीका विभिन्न कलेजहरुमा गरिसकिएको छ । उक्त स्थानहरुमा पोष्टकार्ड मार्फत घुमफिर वर्ष २०७३ को पनि प्रचार प्रसार गरिएको थियो ।<br>किशोर शर्मा र नरेन्द्र मैनालीद्वारा धनकुटा, सिन्धुपाल्चोक र रसुवामा खिचिएको फोटोहरु प्रदर्शनीमा हुनेछन् । |

Figure 5.1: Sample dataset

List of stop-words are given in Figure 5.2.



हौं, हैन, छ, र, पनि, छन्, लागि, भएको, गरेको, भने, गर्न, गर्ने, हो, तथा, यो, रहेको, उनले, थियो, हुने, गरेका, थिए, गर्दै, तर, नै, को, मा, हुन्, भन्ने, हुन, गरी, त, हुन्छ, अब, के, रहेका, गरेर, छैन, दिए, भए, यस, ले, गर्नु, औं, सो, त्यो, कि, जुन, यी, का, गरि, ती, न, छु, छौं, लाई, नि, उप, अक्सर, आदि, कसरी, क्रमशः, चाले, अगाडी, अझै, अनुसार, अन्तर्गत, अन्य, अन्यत्र, अन्यथा, अरु, अरुलाई, अर्को, अर्थात, अर्थात्, अलग, आए, आजको, ओठ, आत्म, आफू, आफूलाई, आफ्नै, आफ्नो, आयो, उदाहरण, उनको, उहालाई, एउटै, एक, एकदम, कतै, कसै, कसैले, कहाँबाट, कहिलेकाहीं, का, किन, किनभने, कुनै, कुरा, कृपया, केही, कोही, गए, गरौं, गर्छ, गर्छु, गर्नुपर्छ, गयौ, गैर, चार, चाहनुहुन्छ, चाहन्छु, चाहिए, छू, जताततै, जब, जबकि, जसको, जसबाट, जसमा, जसलाई, जसले, जस्तै, जस्तो, जस्तोसुकै, जहाँ, जान, जाहिर, जे, जो, ठीक, तत्काल, तदनुसार, तपाईको, तपाई, पर्याप्त, पहिले, पहिलो, पहिल्यै, पाँच, पाँचौं, तल, तापनी, तिनी, तिनीहरू, तिनीहरुको, तिनिहरुलाई, तिमी, तिर, तीन, तुरुन्तै, तेस्रो, तेस्कारण, पूर्व, प्रति, प्रतेक, प्लस, फेरी, बने, त्सपछि, त्सैले, त्यहाँ, थिएन, दिनुभएको, दिनुहुन्छ, दुई, देखि, बरु, बारे, बाहिर, देखिन्छ, देखियो, देखे, देखेको, देखेर, दोस्रो, धेरै, नजिकै, नत्र, नयाँ, निम्ति, बाहेक, बीच, बीचमा, भन, निम्न, निम्नानुसार, निर्दिष्ट, नौ, पक्का, पक्कै, पछि, पछिल्लो, पटक, पर्छ, पर्थ्यो, भन्छन्, भन्, भन्छु, भन्दा, भन्नुभयो, भयो, भर, भित्र, भित्री, म, मलाई, मात्र, माथि, मुख्य, मेरो, यति, यथोचित, यदि, यद्यपि, यसको, यसपछि, यसबाहेक, यसरी, यसो, यस्तो, यहाँ, यहाँसम्म, या, रही, राखे, राख्छ, राम्रो, रूप, लगभग, वरीपरी, वास्तवमा, बिरुद्ध, बिशेष, सायद, शायद, संग, संगै, सक्छ, सट्टा, सधै, सबै, सबैलाई, समय, सम्भव, सम्म, सही, साँच्चै, सात, साथ, साथै, सारा, सोही, स्पष्ट, हरे, हरेक

Figure 5.2: Stopwords

# Chapter 6

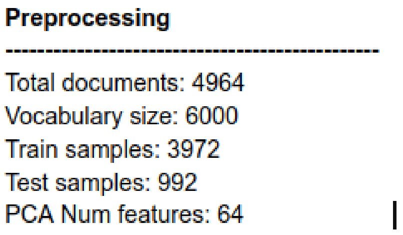
# Experimental Results

Implemented clustering approaches are experimentally evaluated the performance of the different clustering metrics on different dataset settings. In the rest of this chapter we describe the various dataset settings, experimental methodology, and experimental results. As an evaluation, higher Dunn index indicates better clustering, lower DB index indicates the better clustering, higher CH index indicates the better clustering, lower clustering time and training time is again good for real-time clustering.

## 6.1 Implementation

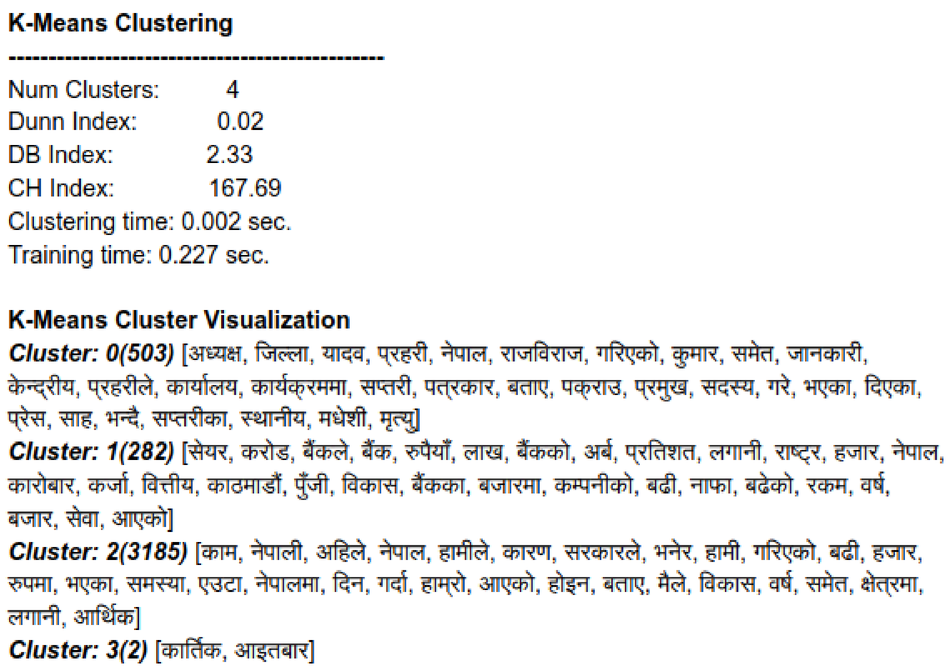
For the implementation of the thesis approaches we used JVM execution environment and Java language. We used smile (https://github.com/haifengl/smile) as a supporting library for the implementation of clustering algorithms.

## 6.2 Clustering samples



```
Preprocessing
----------------------------------------------
Total documents: 4964
Vocabulary size: 6000
Train samples: 3972
Test samples: 992
PCA Num features: 64
```

Figure 6.1: Clustering document pre-processing



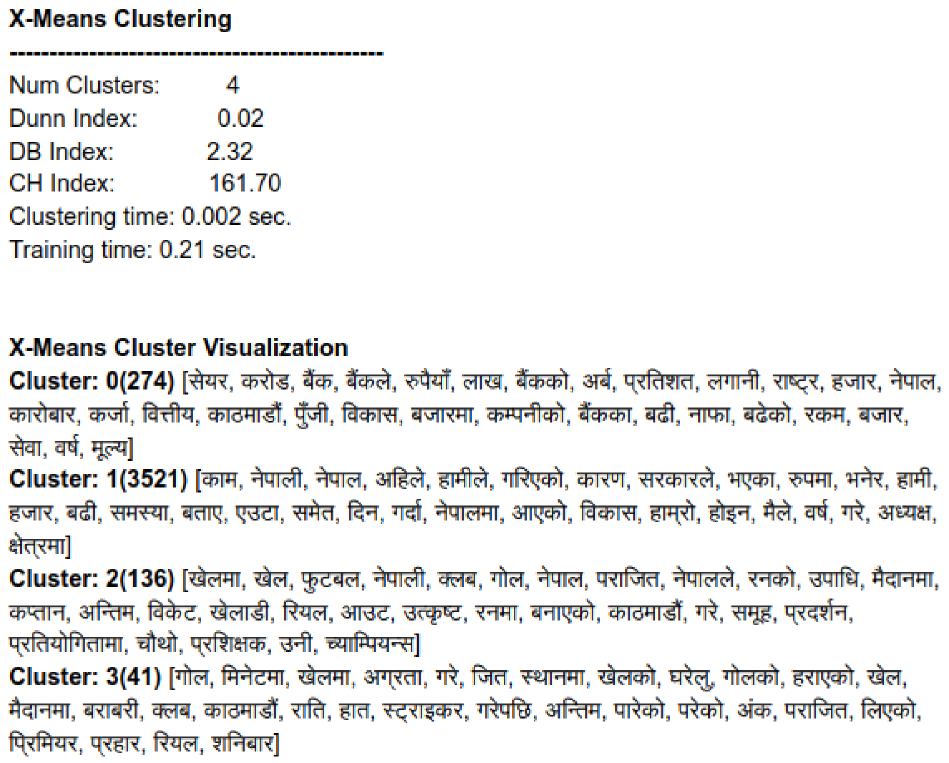
**K-Means Clustering**

```
-----------------------------------------------
Num Clusters:       4
Dunn Index:         0.02
DB Index:           2.33
CH Index:           167.69
Clustering time: 0.002 sec.
Training time: 0.227 sec.
```

**K-Means Cluster Visualization**

***Cluster: 0(503)*** [अध्यक्ष, जिल्ला, यादव, प्रहरी, नेपाल, राजविराज, गरिएको, कुमार, समेत, जानकारी, केन्द्रीय, प्रहरीले, कार्यालय, कार्यक्रममा, सप्तरी, पत्रकार, बताए, पक्राउ, प्रमुख, सदस्य, गरे, भएका, दिएका, प्रेस, साह, भन्दै, सप्तरीका, स्थानीय, मधेशी, मृत्यु]
***Cluster: 1(282)*** [सेयर, करोड, बैंकले, बैंक, रुपैयाँ, लाख, बैंकको, अर्ब, प्रतिशत, लगानी, राष्ट्र, हजार, नेपाल, कारोबार, कर्जा, वित्तीय, काठमाडौं, पुँजी, विकास, बैंकका, बजारमा, कम्पनीको, बढी, नाफा, बढेको, रकम, वर्ष, बजार, सेवा, आएको]
***Cluster: 2(3185)*** [काम, नेपाली, अहिले, नेपाल, हामीले, कारण, सरकारले, भनेर, हामी, गरिएको, बढी, हजार, रुपमा, भएका, समस्या, एउटा, नेपालमा, दिन, गर्दा, हाम्रो, आएको, होइन, बताए, मैले, विकास, वर्ष, समेत, क्षेत्रमा, लगानी, आर्थिक]
***Cluster: 3(2)*** [कार्तिक, आइतबार]

Figure 6.2: K-Means Clustering Sample



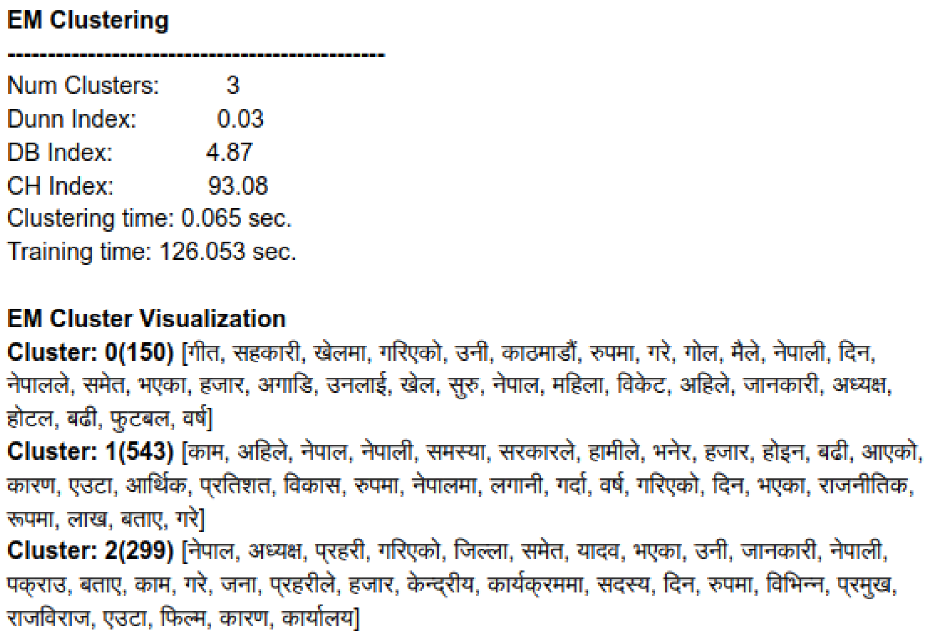
**X-Means Clustering**

```
-----------------------------------------------
Num Clusters:       4
Dunn Index:         0.02
DB Index:           2.32
CH Index:           161.70
Clustering time: 0.002 sec.
Training time: 0.21 sec.
```

**X-Means Cluster Visualization**

**Cluster: 0(274)** [सेयर, करोड, बैंक, बैंकले, रुपैयाँ, लाख, बैंकको, अर्ब, प्रतिशत, लगानी, राष्ट्र, हजार, नेपाल, कारोबार, कर्जा, वित्तीय, काठमाडौं, पुँजी, विकास, बजारमा, कम्पनीको, बैंकका, बढी, नाफा, बढेको, रकम, बजार, सेवा, वर्ष, मूल्य]
**Cluster: 1(3521)** [काम, नेपाली, नेपाल, अहिले, हामीले, गरिएको, कारण, सरकारले, भएका, रुपमा, भनेर, हामी, हजार, बढी, समस्या, बताए, एउटा, समेत, दिन, गर्दा, नेपालमा, आएको, विकास, हाम्रो, होइन, मैले, वर्ष, गरे, अध्यक्ष, क्षेत्रमा]
**Cluster: 2(136)** [खेलमा, खेल, फुटबल, नेपाली, क्लब, गोल, नेपाल, पराजित, नेपालले, रनको, उपाधि, मैदानमा, कप्तान, अन्तिम, विकेट, खेलाडी, रियल, आउट, उत्कृष्ट, रनमा, बनाएको, काठमाडौं, गरे, समूह, प्रदर्शन, प्रतियोगितामा, चौथो, प्रशिक्षक, उनी, च्याम्पियन्स]
**Cluster: 3(41)** [गोल, मिनेटमा, खेलमा, अग्रता, गरे, जित, स्थानमा, खेलको, घरेलु, गोलको, हराएको, खेल, मैदानमा, बराबरी, क्लब, काठमाडौं, राति, हात, स्ट्राइकर, गरेपछि, अन्तिम, पारेको, परेको, अंक, पराजित, लिएको, प्रिमियर, प्रहार, रियल, शनिबार]

Figure 6.3: K-Means Clustering Sample



```
EM Clustering
---------------------------------------------
Num Clusters:       3
Dunn Index:         0.03
DB Index:          4.87
CH Index:          93.08
Clustering time: 0.065 sec.
Training time: 126.053 sec.

EM Cluster Visualization
Cluster: 0(150) [गीत, सहकारी, खेलमा, गरिएको, उनी, काठमाडौं, रुपमा, गरे, गोल, मैले, नेपाली, दिन,
नेपालले, समेत, भएका, हजार, अगाडि, उनलाई, खेल, सुरु, नेपाल, महिला, विकेट, अहिले, जानकारी, अध्यक्ष,
होटल, बढी, फुटबल, वर्ष]
Cluster: 1(543) [काम, अहिले, नेपाल, नेपाली, समस्या, सरकारले, हामीले, भनेर, हजार, होइन, बढी, आएको,
कारण, एउटा, आर्थिक, प्रतिशत, विकास, रुपमा, नेपालमा, लगानी, गर्दा, वर्ष, गरिएको, दिन, भएका, राजनीतिक,
रूपमा, लाख, बताए, गरे]
Cluster: 2(299) [नेपाल, अध्यक्ष, प्रहरी, गरिएको, जिल्ला, समेत, यादव, भएका, उनी, जानकारी, नेपाली,
पक्राउ, बताए, काम, गरे, जना, प्रहरीले, हजार, केन्द्रीय, कार्यक्रममा, सदस्य, दिन, रुपमा, विभिन्न, प्रमुख,
राजविराज, एउटा, फिल्म, कारण, कार्यालय]
```

Figure 6.4: K-Means Clustering Sample

## 6.3 K-Means Clustering Experiments

Table 6.2 shows the experimental results for K-Means clustering algorithm. The dataset and experimental setup for K-means algorithm experiment is given below.

Table 6.1: K-Means clustering setup

| | |
|---|---|
| Training data size | 3972 |
| Test data size | 992 |
| Vocabulary size | 10000 |

Table 6.2: Experimentation Report of K Means

| Clusters # | Dunn Index | DB Index | CH Index | Clustering time (sec.) | Training time (min.) |
|---|---|---|---|---|---|
| 3 | 0.27 | 0.58 | 2.96 | 0.123 | 0.67 |
| 4 | 0.2 | 1.05 | 3.94 | 0.163 | 0.689 |
| 5 | 0.36 | 1.26 | 0.62 | 0.05 | 0.689 |
| 8 | 0.08 | 2.92 | 6.54 | 0.338 | 0.907 |
| 10 | 0.1 | 2.85 | 4.54 | 0.43 | 0.902 |
| 12 | 0.09 | 4.14 | 11.03 | 0.481 | 1.18 |
| 14 | 0.08 | 2.23 | 3.23 | 0.56 | 0.995 |
| 16 | 0.15 | 1.73 | 4.96 | 0.632 | 0.937 |
| 18 | 0.08 | 2.91 | 5.1 | 0.714 | 1.34 |
| 20 | 0.08 | 2.94 | 5.58 | 0.789 | 1.41 |

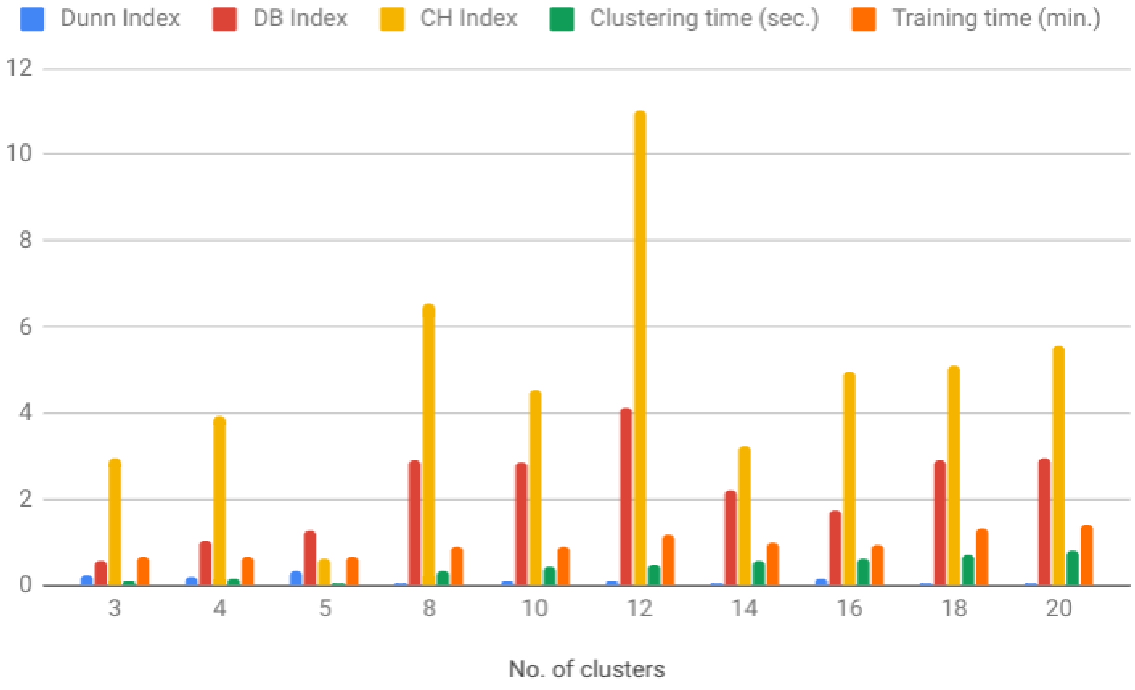


Figure 6.5: K-Means Results

### 6.4 X-Means Clustering Experiments

X-Means clustering algorithm experiments are done with various dataset and experimental configurations. Table 6.4 shows the experimental results for X-means algorithm. Bellow is given the dataset spit for train and test.

Table 6.3: X-Means clustering setup

| | |
|---|---|
| Training data size | 3972 |
| Test data size | 992 |
| Vocabulary size | 10000 |

Table 6.4: Experimentation Report of X Means

| **Clusters #** | **Dunn Index** | **DB Index** | **CH Index** | **Clustering time (sec.)** | **Training time (min.)** |
|---|---|---|---|---|---|
| 2 | 0.34 | 2.69 | 14.62 | 2.86 | 0.43 |
| 4 | 0.16 | 4.89 | 7.79 | 0.017 | 0.688 |
| 6 | 0.22 | 4.39 | 5.51 | 0.025 | 1.05 |
| 8 | 0.05 | 6.53 | 5.64 | 0.033 | 1.29 |
| 10 | 0.05 | 5.67 | 3.24 | 0.04 | 1.51 |
| 12 | 0.05 | 5.21 | 3.71 | 0.049 | 1.62 |
| 14 | 0.05 | 4.3 | 2.97 | 0.055 | 2.09 |
| 16 | 0.15 | 4.73 | 3.7 | 0.064 | 2.87 |
| 18 | 0.15 | 4.9 | 3.52 | 0.071 | 2.87 |
| 20 | 0.05 | 4.25 | 3.17 | 0.077 | 2.46 |

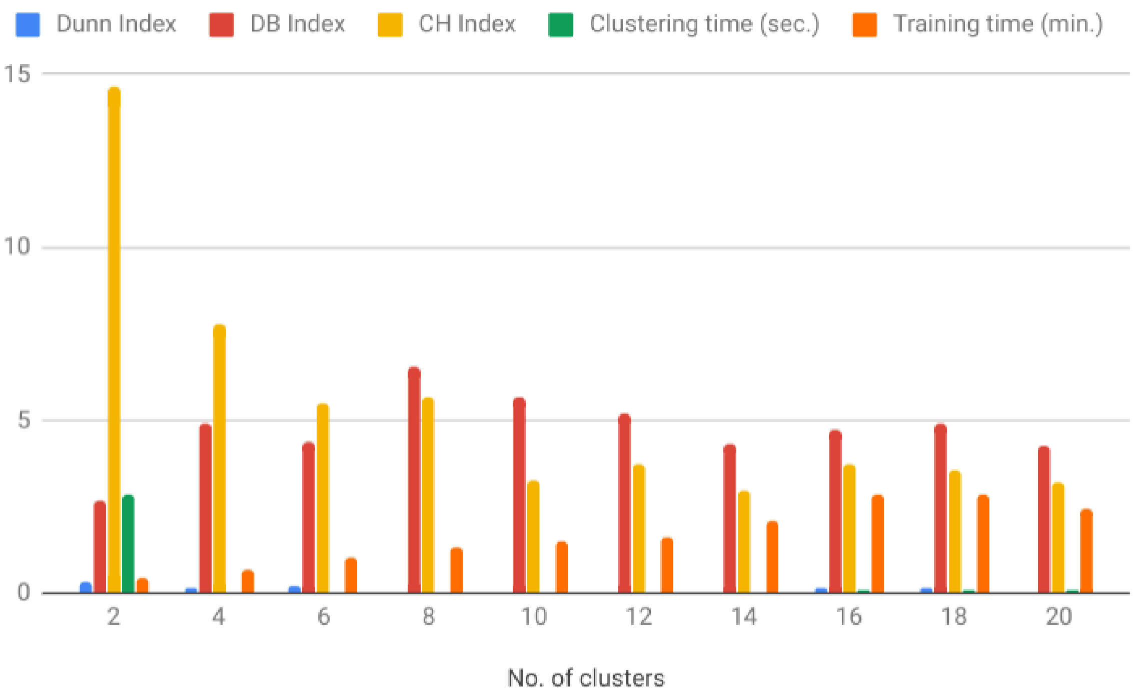


Figure 6.6: X-Means Results

## 6.5 EM Clustering Experiments

EM clustering experiments are done with reduced vocabulary size with PCA. EM algorithms seems computationally complex as it is given a large feature dimension. So, PCA is used for conditionality reduction of large feature space to smaller one.

EM clustering experiments are given in Table 6.6. Dataset spit for the experiment is given below,

Table 6.5: EM clustering setup

| | |
|---|---|
| Training data size | 3972 |
| Test data size | 992 |
| Vocabulary size | 4000 |

Table 6.6: Experimentation Report of EM

| Clusters # | PCA # | Dunn Index | DB Index | CH Index | Clust. time (sec.) | Train time (min.) |
|---|---|---|---|---|---|---|
| 3 | 10 | 0.01 | 2.55 | 314.68 | 0.008 | 0.037 |
| 4 | 32 | 0.02 | 3.29 | 157.72 | 0.042 | 0.34 |
| 3 | 64 | 0.03 | 4.80 | 93.29 | 0.196 | 1.08 |

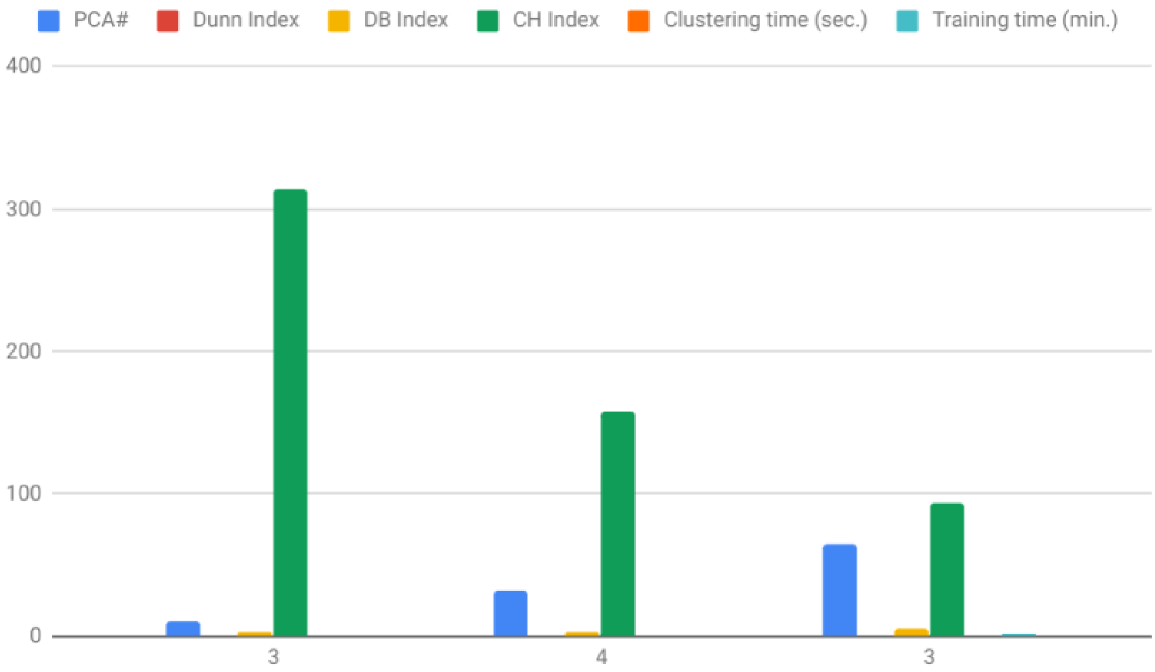


Figure 6.7: EM Results

### 6.6 Experiments with dimensionality Reduction

To reduce the feature embedding dimensions PCA is applied to original feature dimension. Reduced feature dimension gives computationally efficient and accurate performance.

#### 6.6.1 Experiment1

Experiment1 setup is given below.

Table 6.7: Setup1

| | |
|---|---|
| Training data size | 3972 |
| Test data size | 992 |
| Vocabulary size | 4000 |
| PCA | 64 |

K-means clustering experiments for the setup given in Table 6.7 is given in the Table 6.8.

Table 6.8: Experimentation Report of K-Means using setup Table 6.7

| **Clusters #** | **Dunn Index** | **DB Index** | **CH Index** | **Clustering time (sec.)** | **Training time(min.)** |
|---|---|---|---|---|---|
| 2 | 0.04 | 2.52 | 42.29 | 0.001 | 0.399 |
| 4 | 0.14 | 1.33 | 47.16 | 0 | 0.07 |
| 6 | 0.08 | 2.19 | 42.01 | 0.002 | 0.118 |
| 8 | 0.07 | 1.76 | 42.94 | 0.001 | 0.089 |
| 10 | 0.04 | 2.27 | 44.16 | 0.001 | 0.18 |
| 12 | 0.09 | 1.66 | 38.78 | 0.003 | 0.116 |
| 14 | 0.08 | 1.28 | 29.57 | 0.001 | 0.223 |
| 16 | 0.03 | 1.82 | 35.21 | 0.004 | 0.522 |
| 18 | 0.09 | 1.56 | 28.2 | 0.001 | 0.403 |
| 20 | 0.05 | 1.77 | 30.62 | 0.001 | 0.216 |

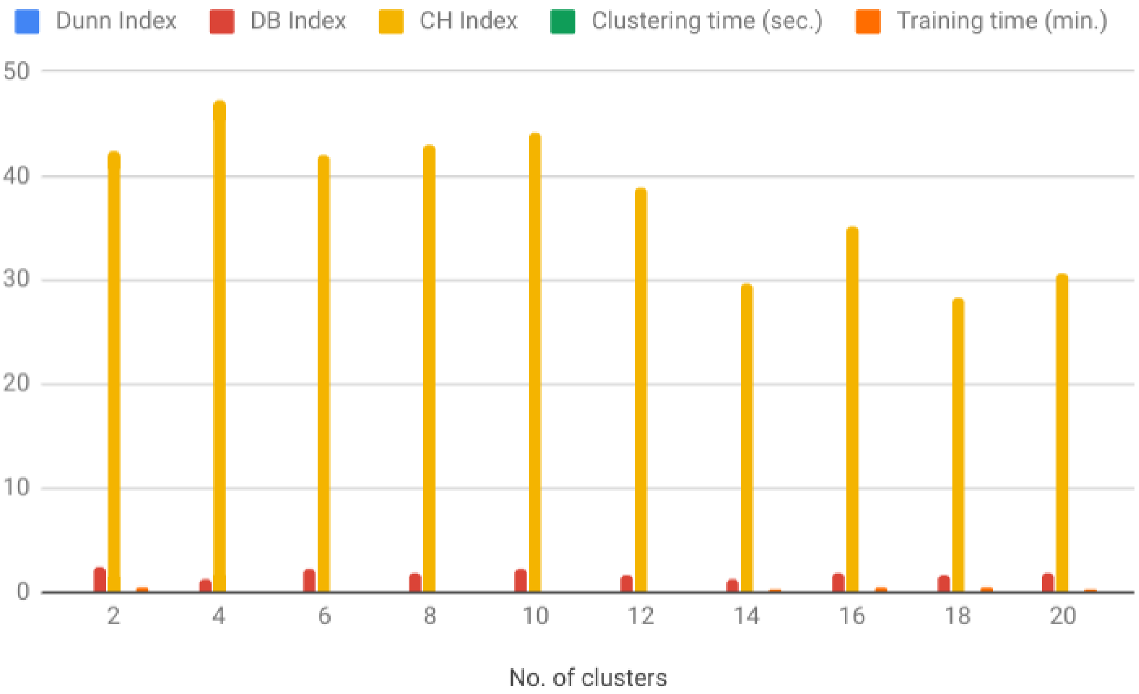


Figure 6.8: K-Means Experiment1 Results

X-means clustering experiments for the setup given in Table 6.7 is given in the Table 6.9.

Table 6.9: Experimentation Report of X-Means using setup Table 6.7

| Clusters # | Dunn Index | DB Index | CH Index | Clustering time (sec.) | Training time (min.) |
|---|---|---|---|---|---|
| 2 | 0.32 | 1.21 | 90.54 | 0 | 0.105 |
| 4 | 0.09 | 2.38 | 61.03 | 0 | 0.21 |
| 6 | 0.09 | 1.91 | 38.4 | 0.001 | 0.218 |
| 8 | 0.08 | 2.02 | 43.98 | 0.001 | 0.283 |
| 10 | 0.09 | 1.66 | 39.6 | 0 | 0.304 |
| 12 | 0.04 | 1.93 | 38.43 | 0.001 | 0.41 |
| 14 | 0.08 | 1.48 | 32.5 | 0.001 | 0.383 |
| 16 | 0.05 | 1.8 | 36.48 | 0.001 | 0.635 |
| 18 | 0.07 | 1.79 | 35.37 | 0.001 | 0.861 |
| 20 | 0.06 | 1.65 | 32.53 | 0.002 | 0.813 |

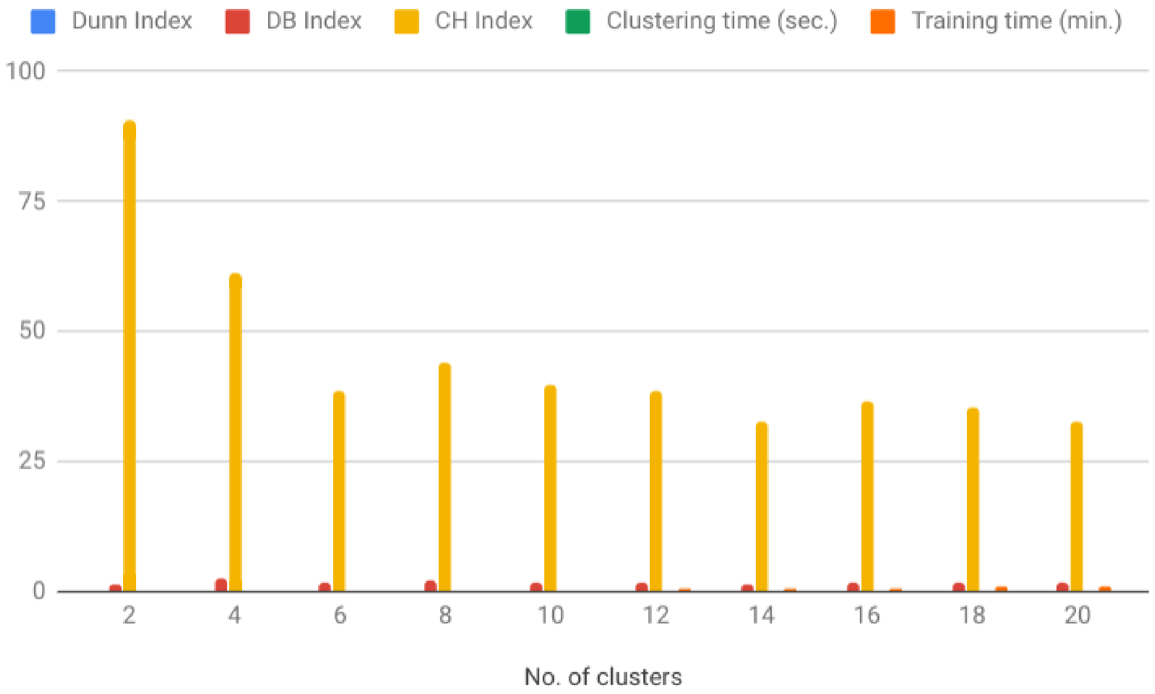


Figure 6.9: X-Means Experiment1 Results

EM clustering experiments for the setup given in Table 6.7 is given in the Table 6.10.

Table 6.10: Experimentation Report of EM using setup Table 6.7

| Clusters # | Dunn Index | DB Index | CH Index | Clustering time (sec.) | Training time (min.) |
|---|---|---|---|---|---|
| 3 | 0.04 | 4.47 | 29.55 | 0.014 | 0.89 |

### 6.6.2 Experiment2

Experiment2 setup is given below.

Table 6.11: Setup2

| | |
|---|---|
| Training data size | 3972 |
| Test data size | 992 |
| Vocabulary size | 4000 |
| PCA | 128 |

K-means clustering experiments for the setup given in Table 6.11 is given in the Table 6.12.

Table 6.12: Experimentation Report of K-Means using setup Table 6.11

| Clusters # | Dunn Index | DB Index | CH Index | Clustering time (sec.) | Training time(min.) |
|---|---|---|---|---|---|
| 2 | 0.14 | 2.12 | 15.21 | 0.001 | 0.293 |
| 4 | 0.1 | 1.9 | 38.86 | 0.001 | 0.098 |
| 6 | 0.09 | 0.78 | 19.83 | 0.001 | 0.063 |
| 8 | 0.02 | 3 | 30.2 | 0.001 | 0.369 |
| 10 | 0.02 | 2.53 | 25.15 | 0.001 | 0.514 |
| 12 | 0.02 | 1.89 | 22.83 | 0.001 | 0.244 |
| 14 | 0.08 | 1.28 | 20.59 | 0.002 | 0.16 |
| 16 | 0.03 | 2.11 | 32.69 | 0.002 | 1.369 |
| 18 | 0.02 | 1.93 | 18.24 | 0.003 | 0.576 |
| 20 | 0.08 | 1.83 | 27.07 | 0.002 | 0.923 |

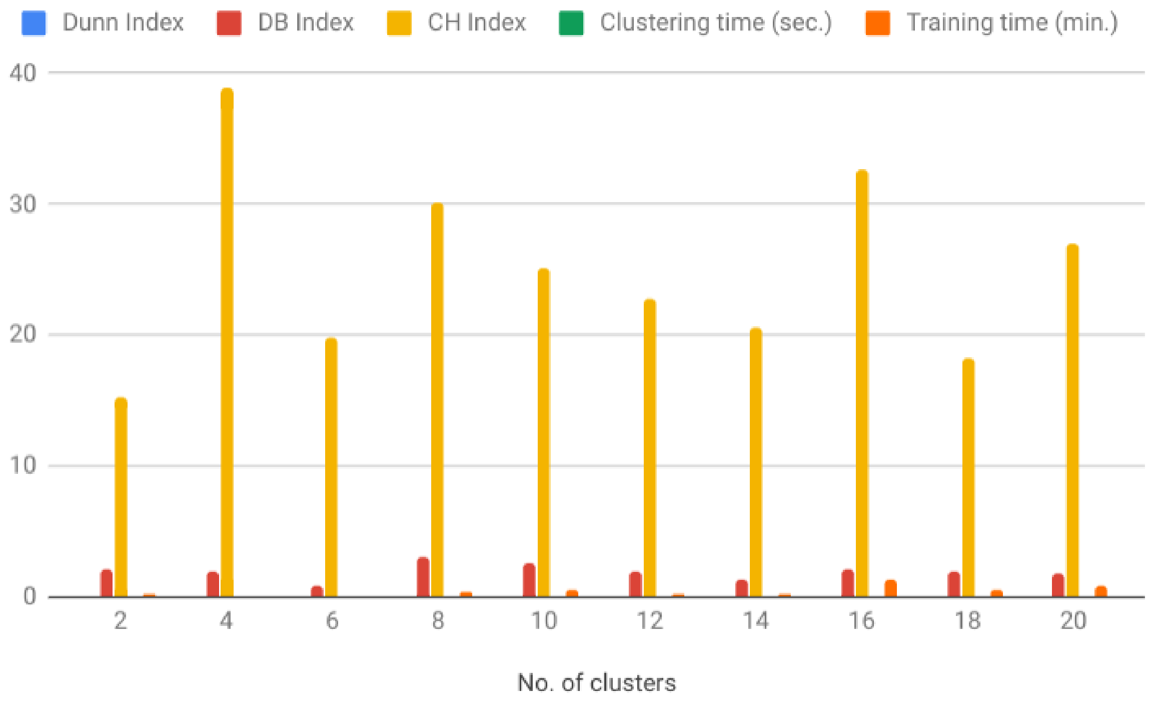


Figure 6.10: K-Means Experiment2 Results

X-means clustering experiments for the setup given in Table 6.11 is given in the Table 6.13.

Table 6.13: Experimentation Report of X-Means using setup Table 6.11

| Clusters # | Dunn Index | DB Index | CH Index | Clustering time (sec.) | Training time (min.) |
|---|---|---|---|---|---|
| 2 | 0.1 | 1.28 | 75.48 | 0.001 | 0.187 |
| 4 | 0.06 | 2.35 | 43.42 | 0.001 | 0.249 |
| 6 | 0.02 | 2.49 | 36.01 | 0.001 | 0.494 |
| 8 | 0.12 | 2.06 | 46.92 | 0.001 | 0.436 |
| 10 | 0.02 | 2.74 | 28.62 | 0.001 | 0.775 |
| 12 | 0.06 | 2.44 | 33.36 | 0.001 | 0.732 |
| 14 | 0.06 | 2.1 | 30.14 | 0.002 | 0.851 |
| 16 | 0.003 | 2.26 | 26.94 | 0.001 | 1.477 |
| 18 | 0.03 | 2.21 | 21.76 | 0.002 | 1.948 |
| 20 | 0.03 | 2.09 | 24.92 | 0.003 | 1.94 |

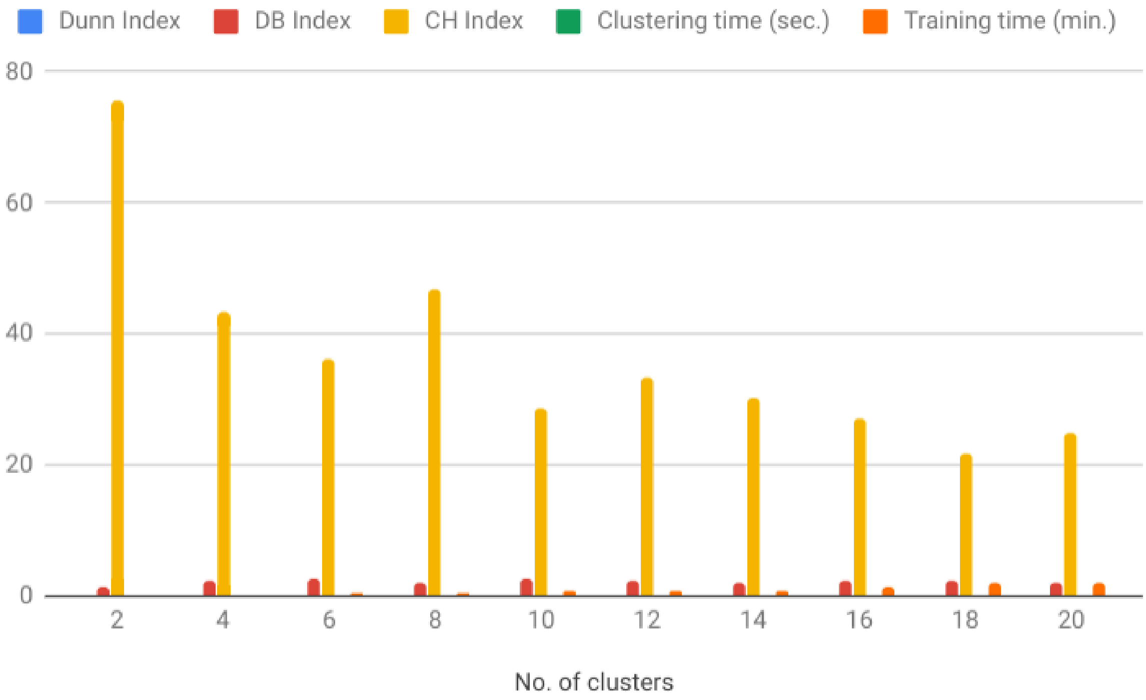


Figure 6.11: X-Means Experiment2 Results

EM clustering experiments for the setup given in Table 6.7 is given in the Table 6.14.

Table 6.14: Experimentation Report of EM using setup Table 6.7

| Clusters # | Dunn Index | DB Index | CH Index | Clustering time (sec.) | Training time (min.) |
|---|---|---|---|---|---|
| 4 | 0.02 | 4.52 | 60.68 | 0.077 | 1.01 |

### 6.6.3 Experiment3

Experiment3 setup is given below.

Table 6.15: Setup3

| | |
|---|---|
| Training data size | 3972 |
| Test data size | 992 |
| Vocabulary size | 6000 |
| PCA | 64 |

K-means clustering experiments for the setup given in Table 6.15 is given in the Table 6.16.

Table 6.16: Experimentation Report of K-Means using setup Table 6.15

| Clusters # | Dunn Index | DB Index | CH Index | Clustering time (sec.) | Training time(min.) |
|---|---|---|---|---|---|
| 2 | 0.03 | 2.94 | 36.31 | 0 | 0.419 |
| 4 | 0.02 | 2.06 | 51.18 | 0 | 0.081 |
| 6 | 0.02 | 1.54 | 41.97 | 0.001 | 0.113 |
| 8 | 0.02 | 1.88 | 42.31 | 0.001 | 0.094 |
| 10 | 0.03 | 2 | 47.78 | 0.001 | 0.25 |
| 12 | 0.02 | 1.63 | 29.93 | 0.001 | 0.123 |
| 14 | 0.05 | 1.71 | 41.42 | 0 | 0.287 |
| 16 | 0.04 | 1.9 | 57.37 | 0.001 | 0.273 |
| 18 | 0.07 | 1.54 | 47.59 | 0.001 | 0.259 |
| 20 | 0.06 | 1.52 | 47.48 | 0.001 | 0.485 |

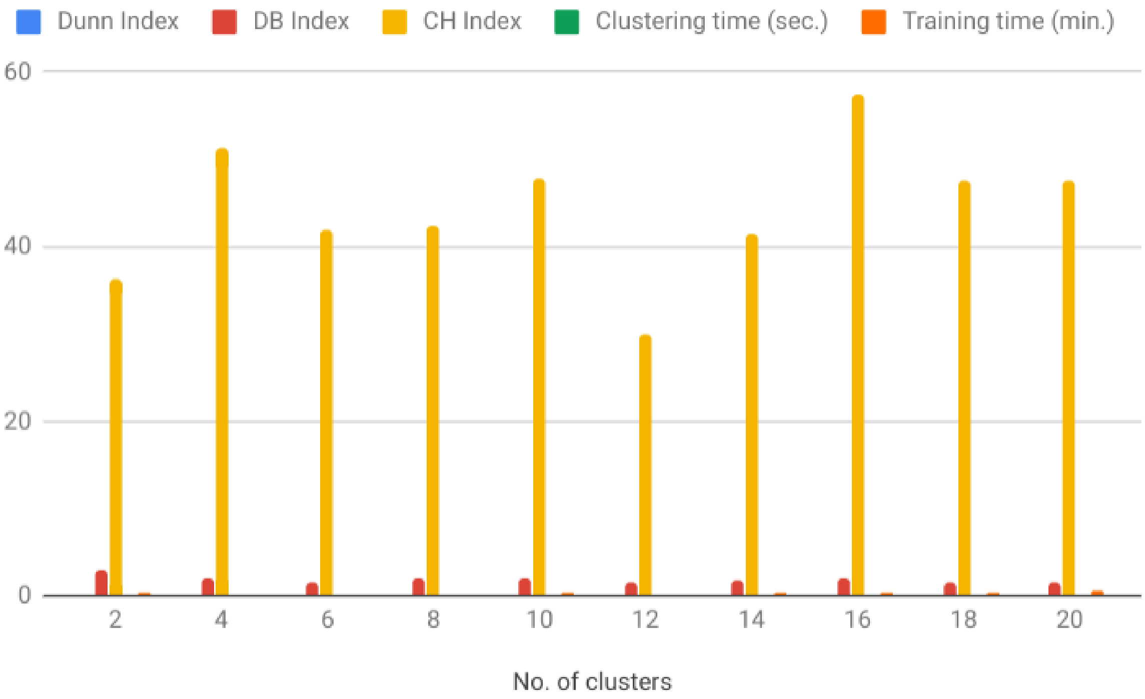


Figure 6.12: K-Means Experiment3 Results

X-means clustering experiments for the setup given in Table 6.15 is given in the Table 6.17.

Table 6.17: Experimentation Report of X-Means using setup Table 6.15

| Clusters # | Dunn Index | DB Index | CH Index | Clustering time (sec.) | Training time (min.) |
|---|---|---|---|---|---|
| 2 | 0.07 | 1.38 | 81.77 | 0 | 0.583 |
| 4 | 0.03 | 2.3 | 53.92 | 0 | 0.116 |
| 6 | 0.02 | 2.06 | 46.39 | 0 | 0.186 |
| 8 | 0.02 | 1.85 | 37.99 | 0 | 0.18 |
| 10 | 0.05 | 2.12 | 60.02 | 0 | 0.287 |
| 12 | 0.01 | 1.98 | 34.67 | 0.001 | 0.42 |
| 14 | 0.04 | 1.55 | 48.85 | 0.01 | 0.724 |
| 16 | 0.02 | 1.57 | 37.38 | 0.001 | 1.061 |
| 18 | 0.02 | 1.72 | 45.44 | 0.001 | 0.648 |
| 20 | 0.02 | 2.11 | 31.31 | 0.001 | 1.024 |

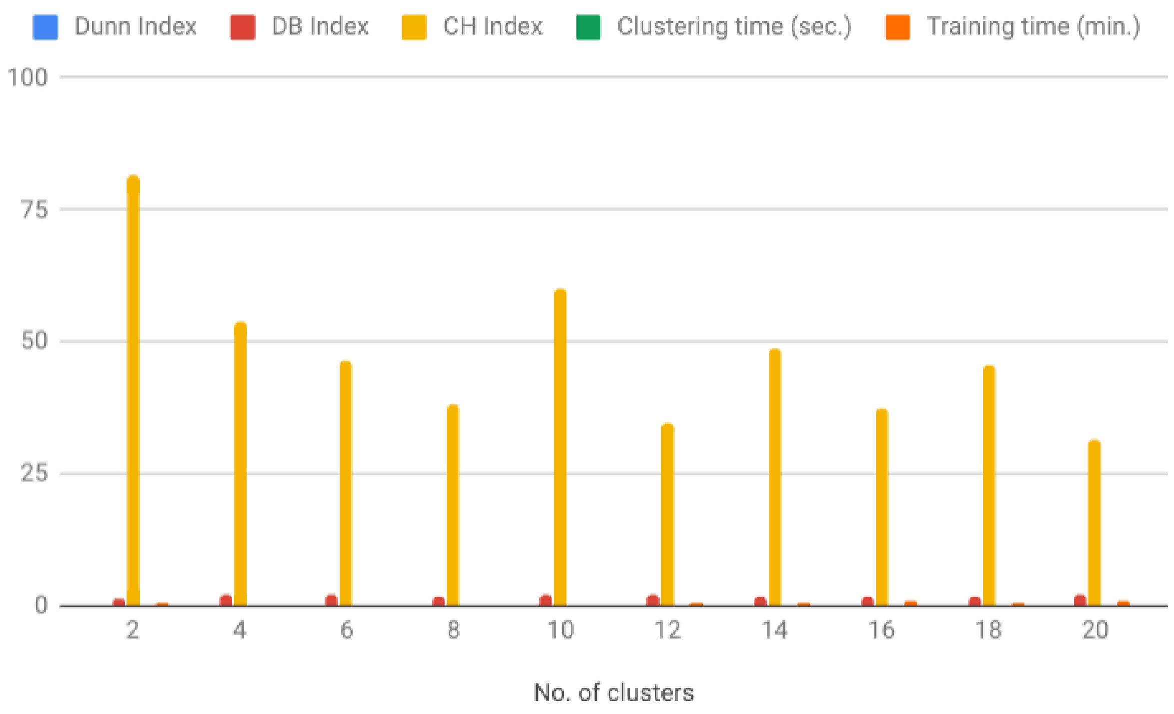


Figure 6.13: X-Means Experiment3 Results

EM clustering experiments for the setup given in Table 6.15 is given in the Table 6.18.

Table 6.18: Experimentation Report of EM using setup Table 6.15

| Clusters # | Dunn Index | DB Index | CH Index | Clustering time (sec.) | Training time (min.) |
|---|---|---|---|---|---|
| 4 | 0.02 | 4.19 | 101.68 | 0.024 | 1.73 |

# Chapter 7

# Conclusion

## 7.1 Conclusion

Evaluating best clustering algorithms is very crucial and important to the clustering task since, there are number of clustering algorithms known till now. Three different algorithms for clustering evaluation of Nepali news is discussed in this research work. K-Means, X means and EM algorithms are compared with each other with the help of Dunn index, DB index, CH index,clustering time and training time. When using those algorithms for clustering large feature size of the dataset took much clustering and training time. So, for better computational efficiency and accuracy PCA is used as a dimensionality reduction algorithm.

Experimental results shows for large vocab size and extracted features, X-means clustering algorithm perform better then K-means and EM. EM computationally takes much higher time for training. With reduced feature dimensions EM is good competitor for X-means. K-means perform well for larger vocab size but not win the X-means in accuracy and performance. It can also be concluded that when applying PCA prior to training to decrease the feature dimension gives better performance in term of computational efficiency and accuracy.

To select the winner algorithm and setting the values DB index, training time and clustering time must be lower and value of CH index and Dunn index must be higher. So, based upon the evaluation results, we conclude the winning algorithm and strategies in some states as follows. When feature dimension is high ($>=10000$) K-Means perform better then others. When applied PCA to reduce feature space, EM algorithm better performs than others. With reduced feature space, K-Means still performs better then X-Means clustering algorithm. For the figurative results, we averaged all the values of evaluation indices and found as follows. For K- means the average values of Dunn, DB and CH indices are 0.079, 1.971, 28.086 respectively. For X-means the average values of Dunn, DB and CH indices are 0.076, 2.650, 33.701 respectively. Similarly, for EM the average values of Dunn, DB and CH indices are 0.0250, 4.190, 95.120 respectively.

## 7.2 Recommendations

This research work can be further extended for better accuracy and efficiency. Below are given some of the dimensions that can be explored with this thesis work.

- Including machine learning based word embedding calculation such as word2vec and fastText for better semantic representation of the news documents.
- Experimenting other clustering algorithms like DBSCAN, Hierarchical clustering or deep learning based clustering.
- Experimenting TF-IDF and other embedding based feature extractions methods.

# References

[1] P. Willett, "Recent trends in hierarchic document clustering: a critical review," *Information processing & management*, vol. 24, no. 5, pp. 577–597, 1988.

[2] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035, Society for Industrial and Applied Mathematics, 2007.

[3] F. Gorunescu, *Data Mining: Concepts, models and techniques*, vol. 12. Springer Science & Business Media, 2011.

[4] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[5] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Annals of Data Science*, vol. 2, no. 2, pp. 165–193, 2015.

[6] C. C. Aggarwal and C. K. Reddy, *Data clustering: algorithms and applications*. CRC press, 2013.

[7] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 267–273, ACM, 2003.

[8] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[9] P. N. Deshmukh Dipali, "comprehensive survey on clustering algorithms and similarity measures," *international journal of advanced research in computer engineering and technology*, vol. 4, 2015.

[10] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[11] S. Bisht and A. Paul, "Document clustering: a review," *International Journal of Computer Applications*, vol. 73, no. 11, 2013.

[12] M. Hughes, I. Li, S. Kotoulas, and T. Suzumura, "Medical text classification using convolutional neural networks," *arXiv preprint arXiv:1704.06841*, 2017.

[13] M. Steinbach, G. Karypis, V. Kumar, *et al.*, "A comparison of document clustering techniques," in *KDD workshop on text mining*, vol. 400, pp. 525–526, Boston, 2000.

[14] R. Muscat, "Automatic document clustering using topic analysis," *Technical Report CSAI 2005–01*, pp. 1–16, 2005.

[15] N. Sharma, A. Bajpai, and M. R. Litoriya, "Comparison the various clustering algorithms of weka tools," *facilities*, vol. 4, no. 7, 2012.

[16] R. Agrawal and J. Agrawal, "Analysis of clustering algorithm of weka tool on air pollution dataset," *International Journal of Computer Applications*, vol. 168, no. 13, 2017.

[17] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification.," in *AAAI*, vol. 333, pp. 2267–2273, 2015.

[18] K. Kowsari, D. E. Brown, M. Heidarysafa, K. J. Meimandi, M. S. Gerber, and L. E. Barnes, "Hdltex: Hierarchical deep learning for text classification," *arXiv preprint arXiv:1709.08267*, 2017.

[19] Y. G. Jung, M. S. Kang, and J. Heo, "Clustering performance comparison using k-means and expectation maximization algorithms," *Biotechnology & Biotechnological Equipment*, vol. 28, no. sup1, pp. S44–S48, 2014.

[20] D. Pelleg, A. W. Moore, *et al.*, "X-means: Extending k-means with efficient estimation of the number of clusters.," in *ICML*, vol. 1, pp. 727–734, 2000.

[21] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[22] G. Celeux and G. Govaert, "A classification em algorithm for clustering and two stochastic versions," *Computational statistics & Data analysis*, vol. 14, no. 3, pp. 315–332, 1992.

[23] C. Apté, F. Damerau, and S. M. Weiss, "Automated learning of decision rules for text categorization," *ACM Transactions on Information Systems (TOIS)*, vol. 12, no. 3, pp. 233–251, 1994.

[24] M. Sokolova and G. Lapalme, "A systema@articlewillett1988recent, title=Recent trends in hierarchic document clustering: a critical review, author=Willett, Peter, journal=Information processing & management, volume=24, number=5, pages=577–597, year=1988, publisher=Elsevier tic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[25] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," *Machine learning: ECML-98*, pp. 137–142, 1998.

[26] A. K. Tegegnie, A. N. Tarekegn, and T. A. Alemu, "A comparative study of flat and hierarchical classification for amharic news text using svm," *Culture*, vol. 2007, p. 1, 2010.

[27] T. B. Shahi, T. N. Dhamala, and B. Balami, "Support vector machines based part of speech tagging for nepali text," *International Journal of Computer Applications*, vol. 70, no. 24, 2013.

[28] A. Paul, B. S. Purkayastha, and S. Sarkar, "Hidden markov model based part of speech tagging for nepali language," in *Advanced Computing and Communication (ISACC), 2015 International Symposium on*, pp. 149–156, IEEE, 2015.

[29] K. Kafle, D. Sharma, A. Subedi, and A. K. Timalsina, "Improving nepali document classification by neural network," in *Proceedings of IOE Graduate Conference*, pp. 317–322, 2016.

[30] C. Sitaula, "A hybrid algorithm for stemming of nepali text," *Intelligent Information Management*, vol. 5, no. 04, p. 136, 2013.

[31] S. B. Bam and T. B. Shahi, "Named entity recognition for nepali text using support vector machines," *Intelligent Information Management*, vol. 6, no. 02, p. 21, 2014.

[32] A. Dey, A. Paul, and B. S. Purkayastha, "Named entity recognition for nepali language: A semi hybrid approach," *International Journal of Engineering and Innovative Technology (IJEIT) Volume*, vol. 3, pp. 21–25, 2014.

[33] T. B. Shahi and A. Yadav, "Mobile sms spam filtering for nepali text using naïve bayesian and support vector machine," *International Journal of Intelligence Science*, vol. 4, no. 01, p. 24, 2013.

[34] S. Thakur and V. K. Singh, "A lexicon pool augmented naive bayes classifier for nepali text," in *Contemporary Computing (IC3), 2014 Seventh International Conference on*, pp. 542–546, IEEE, 2014.

[35] H. AbuelFutuh, *News Feeds Clustering Research Study*. PhD thesis, Nova Southeastern University, 2015.

[36] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pp. 911–916, IEEE, 2010.

[37] T. B. Shahi and A. K. Pant, "Nepali news classification using naïve bayes, support vector machines and neural networks," in *2018 International Conference on Communication information and Computing Technology (ICCICT)*, pp. 1–5, IEEE, 2018.

[38] N. Shah and S. Mahajan, "Document clustering: a detailed review," *International Journal of Applied Information Systems*, vol. 4, no. 5, pp. 30–38, 2012.

[39] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.

[40] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of cybernetics*, vol. 4, no. 1, pp. 95–104, 1974.

[41] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.