Emphasis on the key factors derived from employee review text using NLP Techniques

Kedarnath Goud Kothinti

Thesis Report

January 2020

# Acknowledgement

I would like to thank number of people who supported, inspired, motivated and challenged me through this journey. First in the list must start with thanking my advisor Rupal Bhargava, who taught me how to do research, find my path in thesis with on time support as needed. She was also a good mentor, helping me stay motivated and focused throughout this process.

Would like thank Rajesh Jagtap the student mentor with prompt and timely updates to help not missing the deadline dates for submissions.

I would like to thank my family, my mother Saraswathi, my father Sadasivudu, my wife Navya and son Naithik for all their support and encouragement during this course tenure. Finally, I would like to thank Harsha my friend who has motivated and provided confidence to complete this endeavour with continued thoughtful ideas and suggestions.

# Abstract

This thesis proposes an NLP technique to emphasize on the key parameters that can be derived from the employee review text. The goal of our process is to obtain most relevant attributes (text clusters) defined from lucid review text look up for organization with respective to management, work culture, pay structure and more facts. In this thesis we have used word embeddings to create vector representation of word in the corpus by using tWord2Vec, Tf-IDF and Truncated SVD on text before we cluster the review texts. Then we will use k-means, an unsupervised algorithm to make inferences from datasets using only input vectors without referring to known or labelled clusters and sensitive to initial cluster centres. As part of this initial K-Means clustering we will optimize the sparse matrix of weighted words created by normalize the vector and then create clusters of key factors of text using silhouette scores while tuning the K-means algorithm by calculating best cluster estimation. As the clusters doesn’t give us details of topics, we need to make sense to these cluster contents by making use of topic modelling technique Latent Dirichlet Allocation (LDA) to extract the topics within these clusters. By choosing best clusters based on silhouette score and hyper tuning the parameters of LDA has provided insight to better results in identifying the key features from text phrases to segregate and extract the details within text reviews rather than numeric based models. This approach will be very helpful in recommending job seeker to know about organisation based on his expectation and can further expanded to recommendation model to organisation to know the trends in market to attract right talent that will benefit them. The experiments have proven that this was effective in finding key features with employ review text.

# TABLE of contents

Page

# List of figures v

# LIST OF TABLES vi

CHAPTER 1: INTRODUCTION1

* 1. Motivation1
  2. Research Gaps 2
  3. Objectives 7
  4. Scope of work 11
  5. Thesis Contribution 11

CHAPTER 2: LITERATURE REVIEW12

2.1 Introduction 13

2.2 Literature Review 13

2.2.1 Clustering Text Data16

2.2.2 K-Means Clustering Algorithms17

2.2.3 Cluster Content summarization11

2.3 Discussion 13

CHAPTER 3: RESEARCH METHODOLOGY14

* 1. Introduction14
  2. Employee Review Data 14

3.2.1 Data Selection15

3.2.2 Data Pre-processing16

3.2.3 Data Transformation 16

3.2.3 Data Interpretation 17

* 1. Proposed Methodology 16

3.3.1 Introduction17

3.3.2 Desiderata and Motivation18

3.3.3 Proposed Approach19

3.3.3.1 Design Overview19

3.3.3.2 Word Embeddings19

3.3.3.3 K-means clustering to compute optimal clusters19

3.3.3.4 Latent Semantic Analysis using TF-IDF19

3.3.3.5 Apply K-means algorithm on Truncated SVD19

3.3.3.6 Topic Modelling using LDA19

3.3.3.7 LDA hyperparameter Tuning19

3.3.4 Topic visualisation within cluster19

* 1. Summary19

CHAPTER 4: ANALYSIS 15

4.1 Introduction 15

4.2 Dataset Description15

4.3 Data Set Preparation15

4.4 Exploratory Data Analysis 15

4.5 Data Visualization 15

4.6 Summary15

CHAPTER 5: RESULTS AND DISCUSSIONS 15

5.1 Introduction 15

5.2 Interpretation of Visualizations 15

5.3 Evaluation of Results15

5.4 Summary15

CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS 17

6.1 Introduction 15

6.2 Discussion and Conclusion 15

6.3 Contribution to knowledge 15

REFERENCES 154

APPENDIX A: RESEARCH PROPOSAL 167

APPENDIX B: ETHICS FORMS 179

APPENDIX C: GITHUB CODE 179

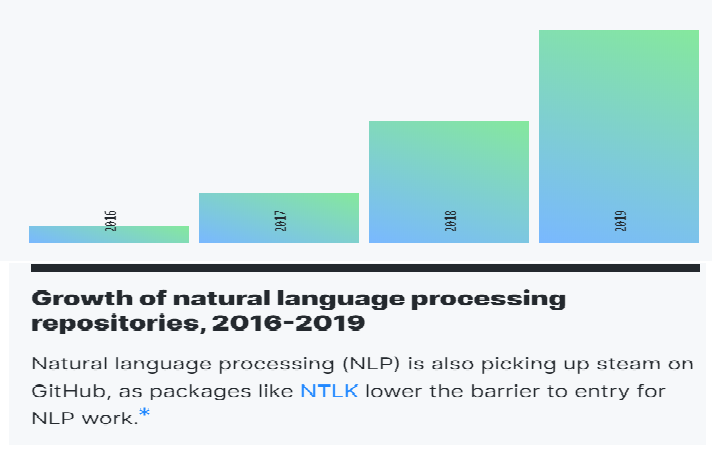
# Chapter 1

**INTRODUCTION**

## Motivation

As per reports 83% increase in the number of people using internet in last 5 years, resulted digital growth since 2016 that contributed to 90% of data has been created s from various sectors like Banking sector, Health sector, Automobile industry, Agriculture, Education sector and others.90% of data created is unstructured- implies that it is typically text heavy and does not follow a predefined data model[OR.1]. Hence storing or maintaining this data for market trends or predictions became prominent in text analytic field with boosting Hadoop and Machine learning platforms to process this data to find meaningful insights for text data gathered from different sources like web, product review internal/external sources, chatbots, speech analysis and others.

With advancement in data storage and processing technology led to Knowledge Discovery from Data (KDD) gained prominence drastically in process of extracting useful patterns from text data. Below graph [1.1] shows the significance growth of Natural Language processing (NLP) research and need for next generation where data driven predictions, analysis, grouping, classification techniques take big lift in technology field.



*1.1 – Significant growth of NLP repositories on GitHub [OR.2]*

As per my banking/IT domain experience and futuristic trends, I had developed interest to gain in-depth knowledge in NLP techniques and to work on model to uncover patterns, trends, sentiments of employee reviews collected during annual surveys. The motivation for this research is to apply the NLP Techniques and improvise - Document clustering algorithms attempt to group similar reviews together. Clustering such employer review text can help management and employees to make their choice clear during hiring process. One of the

central questions asked in this dissertation is:

***To formulate unsupervised learning algorithm that will emphasis on factors like the management, work culture, pay structure for the build/re-build organization based on employee review text.***

In preliminary model results showed us the task of partitioning the dataset into groups, called clusters. The goal is to split up the data in such a way that points within single cluster are very similar and points in different clusters are different.

But these clusters are having topics within these clusters which has more than one feature topic. Hence, we have used Latent Dirichlet Allocation (LDA), topic model that solves this problem to distinguish features within cluster. That why this is very important when attempting to describe the contents of a cluster to the user in a lucid manner. This leads us to the second major set of questions of this dissertation:

* ***How the working hours/workload is affecting employee person life?***
* ***Is Pay Structure a key factor to attract skilled employees to make better organization?***

The research approach is ingenious and will lead into more facts extracted from the employee review text data rather than existing numeric metric-based reviews.

This research will add new dimension of the text review for more benefit of organizations or employees.

## Research Gaps

Unstructured data could be anything: media, imaging, audio, sensor data, text data, and much more. K-Means clustering is one the common and effective algorithm that has proven to be successful with grouping the text data. K-means clustering as the name itself suggests, is a clustering algorithm, with no pre-determined labels defined.

The objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset. You’ll define a target number k, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the centre of the cluster. Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares.

In other words, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The ‘means’ in the K-means refers to averaging of the data; that is, finding the centroid.

***How the K-means algorithm works***

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

It halts creating and optimizing clusters when either:

• The centroids have stabilized — there is no change in their values because the clustering has been successful.

• The defined number of iterations has been achieved.

There is a lot of ongoing work in research world related to NLP on how to overcome the drawbacks of K-means clustering algorithm with combination of feature modelling techniques to extract patterns and insights of unstructured text data.

Below are few research methods and limitations that were captured while this thesis was being worked on:

H. S. Park [H. S. Park,2006], amongst others have come with experiment clustering with medoids rather than means which is new flavour for K-medoids clustering algorithm.

The algorithm converges in less iteration as it calculates the distance matrix once and reuse it for determining new medoids in every step. As the algorithm use same distance matrix the performance of the model is very good in comparison with K-means. The crucial limitation of this algorithm not being used is due distance matrix is calculated once and which will vary for sparse text data which has less words with repeating words leading to outliers and noisy results.

Xinwu[Xinwu, Li ,2012] , amongst others have proposed new text clustering algorithm is presented based on K-means and Self-Organizing Model (SOM) that will isolate point text to improve the drawback of K-means as its sensitive to initial cluster center .The cluster determined by this algorithm is more precise than traditional random choice centers. The limitation of this algorithm is that we need have better domain knowledge to isolate point text to build cluster centers and we cannot generalise this to all fields without prior knowledge.

C. Xiong [C. Xiong, 2017], amongst others have proposed an improved K-means text clustering algorithm by optimizing initial cluster centers. This algorithm intent is similar to previous one where the isolated cluster center is key focus to improve the traditional K-Means. This approach as based on defining the cluster centers by choosing the high-density dataset with both dataset object distance being largest. The results of this algorithm have shown better text clustering precision, but time complexity involved in finding initial cluster centers is more.

Chadha [A. Chadha, 2014], amongst others have come with relative K-means algorithm which does not require cluster number K upfront. This approach they initially create two clusters with large dissimilar sets and then keep comparing next sequence with then if match update density of the existing cluster else based on outliers form new cluster until we exhaust through the data set. The complex statistical computation will be necessary when we apply this to unstructured data like text.

S.S Khan [S.S Khan, 2017], amongst others have proposed an algorithm to compute initial cluster centers for K-means clustering. This algorithm is based on two thoughts that some of the relationships are very similar to each other and that is why they have same cluster association regardless to initial cluster centers selection. Also, a specific attribute may provide some knowledge about initial cluster center. The initial cluster centers computed using this methodology are found to be very close to the desired cluster centers, for iterative clustering algorithms. This procedure is applicable to clustering algorithms for continuous data. The experimental results show improved and stable solutions using the proposed algorithm. This algorithm utilizes data found to be very close to the desired cluster centers, for iterative clustering which makes it sensitive to outliers.

D.Reddy [D. Reddy, 2012], along with others have come up with a unique method that selects the initial cluster centers with the help of Voronoi diagram constructed from the given set of data points. The initial cluster centers are precisely chosen from those points which lie on the larger radius Voronoi circles. Based on these circle formations the selection of initial cluster centers are derived from these circles. The proposed method is experimented on various artificial (hand-made) as well as real world data sets of various dimensions. The resulted proved to be better than the traditional K-means. The limitation of this algorithm is the complex computation involved in it.

T.Su[T. Su,2004] along with others have suggested a method that has shown good results conceptually and practically deterministic divisive hierarchical method will improve K-means performance by making faster convergence of clusters with less iterations considering the PCA-Part (principal components analysis partitioning) for initialization. They have used SSE (sum-squared-error) criterion to the largest eigen value of covariance matrices which give largest SSE.As this approach involves more computations based on data set its suggested for model with less time complexity.

Zhou,L [Zhou, L , 2010] , amongst others have come up with combination of clustering and KNN (K- Nearest Neighbours) algorithm. This method will pre-process the train data with clustering algorithm and then use KNN to make active adjustment to choose clusters. This was proven that it will provide symmetric phenomenon and mitigate the incorrect boundary condition in test data. The experimental results have shown significant performance improvement on the final clusters. As this is making immense re-evaluation of the K in each iteration the misclassification of cluster content can be possible issue before we try to adopt.

S. S. Yu [S. S. Yu , 2018], along with others have proposed combination a Tri-level and a bi-layer K- Means algorithm. Their experimental results show that cluster quality was better than traditional K-means which is prone to sensitive outliers and amenable to initial cluster centers. As the data set may vary based on domain tri-level K-means will solve the problem of sensitive outliers and a bi-layer K-means proved be efficient in determining the initial cluster centers. The limitation of this method is primarily depended on K-means and multiple iterations are require improvising the results when we have sparse data.

X. Xu[X. Xu 2018],along with others have proposed a novel sentence similarity calculation method with flavours of syntactic analysis over word embeddings. Sentence similarity calculation was done in two simple steps – first find out the word embeddings to mitigate the word ambiguity problem and then apply dependency parser to process the internal grammatical of the sentence structure. This method has proven to be effective in comparing the similarity of two sentences. As text has many parameters which grow linearly with the documents size. Although estimating these parameters is not impossible, it is computationally very expensive.

## 1.3 **Objectives**

Significant growth of the NLP field has provided new horizons into Data Mining or KDD (Knowledge Discovery from Data) to find Frequent Pattern Mining, Association Rule Analysis, Classification and Clustering.

This will benefit online channels of modern employment for both employee and employer.

***Emphasis on the key factors derived from employee review text using NLP techniques*** will have specific clusters for lucid review look up of organization to solve basic thesis question:

***“What are the key features that can be extracted from the employee review comments like management, work life balance, pay structure and others?”***

## Scope of Study

To propose an approach of unsupervised machine learning algorithm like K-means that will show case the management, work culture, pay structure and other key factors for the build/re-build organization by hiring skilled resources.

As part of this thesis we would like to propose using the LDA topic modelling on the clusters obtained by K-Means clustering algorithm after choosing the best cluster using silhouette scores.

## Thesis Contribution

This thesis contributes to the area of machine learning to emphasise on the NLP techniques we can use by feature engineering the text content to the word vectors using TF-IDF followed with Truncated SVD method to form initial clusters by using K-Means clustering while dealing with unstructured text data. Specifically, it introduces *Topic Modelling on the clusters extracted from K-Means Clustering Algorithm*, a novel technique in the field of NLP to enhance patterns extracted from the employee review text data rather than existing numeric metric-based reviews which might not give the specific details required by user while browsing on web/internal organization analysis.

# Chapter 2

**BACKGROUND**

## 2.1 Introduction

In this chapter we review several topics which are related to this work. We start by

reviewing the techniques that are being used to improve the problem of clustering finds applicability for several tasks. We then use clustering techniques provide a coherent summary of the collection within cluster, which can be used in order to provide summary insights into the overall content of the underlying corpus. Variants of such methods, especially sentence clustering, can also be used for topic modelling.

## 2.2 Literature Review

In the process of this thesis work my area of research started with how to process unstructured text data and group them to retrieve useful information for the users.

Hence the literature review was based on two main themes:

* Explore better techniques to convert text to numerical vector models?
* What is better clustering algorithm for processing unlabelled text?
* How to make use of topic modelling to label features within cluster and tune hyper parameters of LDA to summarize the content?

### 2.2.1 Clustering Text Data

Clustering technique has been core element in text processing as it clustering finds applicability for several tasks like Clustering and Classification, Pattern Recognition, Text Mining.

My research started with looking at various algorithms and related work in evaluating performance of these cluster algorithms to choose the right one to solve the objective of pattern recognition/extraction from employee review text. As per the prior analysis by [Fasheng Liu,2011] Clustering is an important form of data mining. Clustering algorithm can be divided into the following categories: hierarchical clustering, partitioned clustering, density-based algorithm, self-organizing maps algorithm. At the same time, the text clustering problem has its particularity. On one hand, the text vector is a high-dimensional vector, usually thousands or even ten thousand; On the other hand, the text vector is usually sparse vector, so it is difficult for the choice of cluster center. As an unsupervised machine learning method, because of not need to train the process and manual label document at category in advance, clustering has certain flexibility and high automation handling ability. It is become an important mean which pays attention for more and more researchers. Table 1.0 provides glimpse of Performance Comparison of Clustering Algorithms. From this initial cluster survey, I was able figure out that partition clustering will be good option and further deep dive on various ways we can improve to achieve the objective.

Then is the question how this algorithm can be applied in text data in [Gupta, A ,2018] Clustering is a process of organizing data objects into a set of non-disjoint classes known as clusters such that objects in the same cluster are similar to each other and dissimilar to the objects in another cluster. Document clustering is based on the similar approach, that is, documents are organized into meaningful clusters in such a way that documents in the same cluster represent same topic and those in different cluster represent different topic.

Hence it was proved by this paper that applying semantic document-clustering techniques has improved the cluster coherence and performance very much. The drawbacks of traditional document clustering like Synonymy and Polysemy, Ambiguity and High Dimensionality were mitigated by usage of “Semantic Document Clustering”. It focuses on the relation between signifiers like words, phrases and terms. The meaning of semantic is related with the meaning in language. Semantic document clustering concerns with partitioning the documents into clusters in such a way that documents in the same cluster are like each other but documents in different clusters are dissimilar to each other. Table 2.0 describes the merits and demerits various clustering algorithms.

In this journey I came across a classical book by [Charu C,2013] “Data Clustering: Algorithms and Applications” which has provided detailed algorithm theory to better understand the problem statement I have and how to approach it in different way. This key book was my bible for this thesis where I was able to make right choice of clustering technique that will be appropriate to my data set.

A text document can be represented either in the form of binary data, when we use the presence or absence of a word in the document in order to create a binary vector. In such cases, it is possible to directly use a variety of categorical data clustering algorithms.

A more enhanced representation would include weighting methods based on the frequencies of the individual words in the document as well as frequencies of words in an entire collection (e.g., TF-IDF weighting).Word2Vec and Glove are few word embedding techniques which are predominately used in context independent use case to convert words to numeric vector representations.

As the size of the vocabulary increases the vector size is prone increase, resulting in more computational time while clustering large text data. As remedy for such huge vector, we can factorize it by Singular value decomposition (SVD) to reduce the dimensions by using *Low-Rank Approximation.*

Quantitative data clustering algorithms can be used in conjunction with these frequencies in order to determine the most relevant groups of objects in the data.

As per Charu C the distinguishing characteristics of the text representation are as follows:

1. Less vocabulary with more combinations in document….it tough to apply for even more serious when the documents to be clustered are very short (e.g., when clustering sentences or tweets).

2. While the lexicon of a given corpus of documents may be large, the words are typically correlated with one another.

3. The number of words (or non-zero entries) in the different documents may vary widely. Therefore, it is important to normalize the document representations appropriately during the clustering task.

After appropriate rigor ,I have decided to explore more on the K-Means clustering algorithm how it works and can be tailored for my thesis work by citing work of Chadha, A [Chadha, A,2014]researches and studies are going on to address two of the major limitations of K-means algorithm - ***One to select efficiently the initial centroids and second to remove the need of giving the number of clusters required as input to the***

***algorithm*.**

### 2.2.2 K-Means Clustering Algorithms

As part of this thesis to emphasis on key factors that can be extracted from the employee review text, we have citied various scholars work and understood the prominent work done so far in this section. Xiong, C [Xiong, C,2017] noted K-means clustering algorithm is an influential algorithm in data mining. K-means clustering algorithm is an influential algorithm in data mining. The traditional K-means algorithm has sensitivity to the initial cluster centers, leading to the result of clustering depends on the initial centers excessively. In order to overcome this shortcoming, Xiong, C, proposes an improved K-means text clustering algorithm by optimizing initial cluster centers. There are some methods for selecting initial cluster centers, such as random method, multiple sampling method, distance optimization method, density method, quadratic clustering method. They have formulated method which combines the “distance optimization method” and “density method” to determine the better initial cluster centers. The proposed algorithm eliminates the sensitivity of the K-means algorithm to the initial cluster centers and can get better results of text clustering but has more time complexity to calculate the distances.

There is one more method proposed by Zhu M [Zhu, M, 2014] local optima that are common in K-means clustering can then be effectively reduced. In addition, the algorithm can obtain all initial cluster centers simultaneously (instead of one center at a time) during the dynamic adjustment. Originality/value - The authors presented in this paper an efficient algorithm, which can dynamically adjust initial cluster centers that are randomly selected. The adjusted centers are highly representative, i.e. they are distributed among as many samples as possible. As a result, local optima that are common in K-means clustering can be effectively reduced so that the authors can achieve an improved clustering accuracy.

I came across similar paper on local optima concept by S. S. Khan [S. S. Khan, 2017] to compute initial cluster centers for K-means clustering. This algorithm is based on two observations that some of the patterns are very similar to each other and that is why they have same cluster membership irrespective to the choice of initial cluster centers. Also, an individual attribute may provide some information about initial cluster center. The initial cluster centers computed using this methodology are found to be very close to the desired cluster centers, for iterative clustering algorithms. This procedure is applicable to clustering algorithms for continuous data. Since this procedure was better for continuous data, we can consider this further.

We have studied few more methods proposed by D. Reddy novel method that selects the initial cluster centers with the help of Voronoi diagram constructed from the given set of data points. The initial cluster centers are effectively selected from those points which lie on the boundary of higher radius Voronoi circles. As a result, the proposed method automates the selection of the initial cluster centers to supply them for K-means. This concept needs more computation to apply on the text data.

The next puzzle I had to choose the better value of K in K-Means algorithm which can be done using Elbow Method and Silhouette Method.

The Elbow method uses squared error for each point is the square of the distance of the point from its representation i.e. its predicted cluster center using distance metric like the Euclidean or Manhattan. As the text data can have reparative words elbow method can be uncertain.

The Silhouette Method, measures silhouette score how similar a point is to its own cluster compared to other clusters. Hence, we can use this method to select the optimal K value in our K-Means cluster algorithm.

### 2.2.3 Cluster Content Summarization

The Cluster Content summarization is nothing but finding the meaningful phrases or topics within the unlabelled clusters that was formed by initial clustering of data set.We have various topic modelling techniques like Matrix Factorisation, Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) which are primary used to find sentence similarity with given text corpus. Hence in quest of finding better technique to refine the results of clustering algorithm we have further down the path of research on topic modelling with below proposed method by eminent scholars/researchers.

For any recommendation based, document retrieval engines, topic modelling is much needed building block of NLP to extract meaningful insight and generate them based on user necessity .This is very crucial to convert the results of clustering, classification and sentiment analysis model to present results to beneficiary. Hence forming appropriate features is very important.

This preliminary method by Xu, X [Xu, X,2018], along with others proposed Sentences similarity analysis based on word embedding and syntax analysis has provided me base idea and necessity to further form labels for clustered text. In this method they have initially calculated word embedding to express the word using the dependency parser to analyse the internal grammatical sentence structure rather than customary bad of word similarity to form phrases of the sentence. Here they have used cosine distance to express the word similarity to calculate resemblance degree between words.

Further deep diving on this sentence sense finding came across the matrix factorisation which involves a lot of computation to work on text data.

Then thought of the Probabilistic Latent Semantic Analysis (PLSA) requires large number of parameters which grow linearly with the documents. Although estimating these parameters is not impossible, it is computationally very expensive.

LDA (Latent Dirichlet Distribution) is an alternative topic model that solves this problem as it’s a parametric model. The normal distribution is parameterized by only two parameters - the mean and the standard deviation.

Usop, E[Usop, E, 2017] , amongst others has proposed a way to improvise the coherence value between two sentences in corpus by using feature selection process of Parts of Speech(POS) and then applying LDA .In LDA number of topic parameter is hectic process to change the value of topic parameter and validate the coherence score every time. We can use LDA-GA Approach proposed by Hsu, C. I., & Chiu, C in 2017 to use Genetic Algorithm (GA) in discovering optimal weights for LDA topics, but it is supervised mechanism which will be tedious job for using on text data. Twinandilla [Twinandilla,S,2018] along with others have proposed an combination K-means text clustering and topic modelling using LDA, which was able to provide good document summarizations with better results. This was impressive contribution but still suffers if optimal cluster(K) value is not determined and will result is error prone clusters if the initial K means incorrectly groups the documents into topics.

## 2.3 Discussion

After an extensive literature survey, I was able to deep dive into various industry standards and updates in the field of NLP. Would like to propose a unique way to achieve thesis objective by using the existing NLP word embedding algorithm in combination on K-means cluster and then extract topic post LDA modelling on the clusters. Hence by this we can emphasise on the key features with the employee review text.

# Chapter 3

**RESEARCH METHODOLGY**

## Introduction

In this chapter, I would like to provide more details on how to gather data, process the data to clusters based on the numerical representation of word of employee review text, apply K-means clustering on word vectors to determine the optical(K)

Number of clusters, Topic modelling using Latent Dirichlet allocation (LDA) along with hyperparameter tuning.

## Employee Review Data

### Data Selection

Data Selection is key phrase one must think while selecting the project topic, as we need to review the what type of data ,source of data we need to achieve our objectives.

This step involves bringing in relevant data from different sources. Sources can include both external and internal systems and depends on the domain of the problem.

As the intent of this thesis to study the key factors that will impact employee at any organization. Employee reviews will be primary source for such kind of

This process is pre-requisite to the data selection process which need domain subject matter expertise to able to make appropriate choice of data.

The primary source of data can be any organization which would like to know the pulse of individual employees which provided initial qualitative data which has provided intuitive base for the research and type of data needed for this research.

Since this is sensitive data, we have pull data set from Kaggle which has raw test dataset of actual employee review texts at various organizations in market. This data has many details like ID, start-up location, job title, summary of review, pros, cons, advice to management and overall rating.

### Data Pre-processing

We use various techniques to clean the data, depending on the domain of the problem. Regular expressions are extensively used at this stage in identifying anomalies in the data and transforming them as required. This is process is called Data Pre-processing.

As part of pre-processing we can use common clean up steps like remove numbers ,remove symbols, remove punctuations, remove URL, remove extra spaces, remove emoticons and remove words of length 1 or 2 in review text. By this process we will remain with meaningful words in document. This process will only help with extracting text content and eliminating non-alphabetic characters in text.

### Data Transformation

Data Transformation of text data includes techniques such as “removal of stop words", "converting all text to a standard case", "stemming/lemmatization “and “tokenization” of documents of next processing steps.

Stop words – these are commonly used words which can be ignore based on language. For example in English we have common words like a, an, of, in doesn’t hold any significant meaning while processing text. Hence we can remove such stop word to focus on main context on the sentence.

Stemming - is the process of finding the base/root form of the word rather that its morphological combinations. For example programmers, programming will be reduce to program base form.

Lemmatization - is the process of grouping together the different inflected forms of a word so they can be analysed as a single item. For example word madness will be referred as mad.

Tokenization is most basic step to identify the meaning of the document by looking at each word or phrases of the document. Each small unit of this process is called token which will be basis for finding the words that constitute to document. After doing tokenization we can perform various tasks in the text data like Count the number of words in the text or Count the frequency of the word, that is, the number of times a particular word is present.

### Data Interpretation

Data Interpretation is important step to understand dataset and further make sense of the text data that needs to processed. In data interpretation gives the relevance of the columns with in dataset .As pert of this we need to check on blank fields, missing data ,do we need these columns for our methodology can be answered.

The observation on the data set has non text content fields like place, start date ,Id.

Hence data interpretation is critical step before proceed with Exploratory data analysis(EDA) to choose the section of data which is required of our problem statement.

## Proposed Methodology

### Introduction

In this chapter, we introduce proposed NLP technique – a novel approach designed to meet the objective to determine list of key features that are hidden within employee review text. This approach is unique as we cluster documents using K-means clustering and then look up the topics within these clusters by LDA topic modelling technique.

In first stage of this approach we will convert documents to machine readable vectors using word2vec vectorization and then feed this as input to K-means clustering algorithm .Using best silhouette score we will determine the optimal (K) number of clusters .

In second stage we will use Latent Semantic Analysis (LSA) to create document-term matrix using Tf-IDF Vectorizer and then lower the dimensionality of the matrix by Truncated SVD method .Then apply K-means clustering on this vector with K value from first stage. As an output of this stage we will save the extracted cluster text to file or disk.

In third stage we will use these text clusters and label the topics within these grouped documents by using LDA topic modelling technique. After hyper tuning the parameters of LDA algorithm we can emphasis on the key features within text clusters to group the related review topics together.

### Desiderata and Motivation

Significant growth of the text analytic field has provided new horizons into extract meaningful insight of unsupervised text data in various business sectors. Using NLP techniques will benefit online channels of modern employment for both employee and employer of any organization. Emphasis on the key factors derived from employee review text using NLP techniques will have specific clusters for lucid review look up by target audience.

Text clustering was done mostly on news, movie and social media content ,there is need to explore more into employee review comments. Most of the current rating system are numeric in nature. As part of this research would like to purpose novel Natural Language processing(NLP) technique to constellate prime elements of modern employment review market.

Motivation of this thesis is to formulate hybrid NLP technique that will show case the management ,work culture, pay structure for the build/re-build organization by hiring skilled resources. Will be using Accuracy that will measure performance by proportion of the correctly cluster text samples .We can base this approach to build online text retrieval system to find the related topic content and the companies which maintain high standards of employment and employee relationship.

The research approach is ingenious and will lead into more facts extracted from the employee review text data rather than existing numeric metric-based reviews. Hope this work will add new dimension of the review text for more benefit of organizations or employees.

### Proposed Approach

#### Design Overview

With the growth of Language processing algorithms one must choose right combination of techniques and approach to meet the business requirements.

Below is the Design Overview Fig.1.0 on the various processes involved to extract key features from employee review text content.

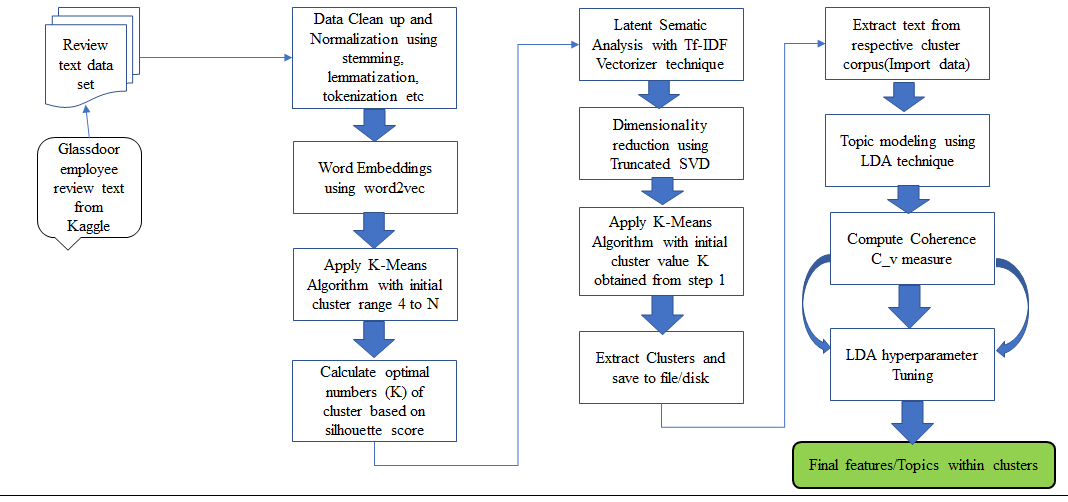


Fig 1.0 Design Overview

#### Word Embeddings

For processing any machine learning algorithm must be represented as machine readable code which can be referred as vector representation.

The process of associating each word to vector in n-dimensional calculated based on how a given word is similar or dissimilar to other in documents is called as *Word Embeddings*.

There different word embedding algorithms available for converting text to vector representations. We can broadly segregate them in to two categories as –

Context Independent embeddings is one of the numeric representation of a word in document irrespective of order they occur or sequence or context with in sentence.Word2Vec and Glove are well know approaches in this category which take input of words and generate word embeddings output without need of any training data set. Here each word can have one word embedding without context of sentence.

Context dependent embeddings each word can have multiple vectors based on the context of the sentence, order of words. Elmo ,BERT and A Lite BERT by Zhenzhong Lan[Zhenzhong Lan,2019] are few algorithms that are more effective as they account for the order of words in sentence. These need to be trained on data set to find various vector representation of word in document.

Based on the our objective and data set interpretation we will be doing Word2Vec embedding technique.

Word2Vec embedding is an efficient statistical way of creating word embedding from given text document. As we are required to run this on each word in document its take like more time to provide word embeddings.Word2Vec provide Continuous Bag-of-Words(CBOW model) and Continuous Skip-Gram Model.

CBOW model – uses context window size for predicting the word embedding while Continuous Skip-Gram Model uses the surrounding sentence in to account for predicting the word embedding.

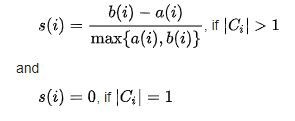
As data set is sparse with limited to employee review domain we can make use of the CBOW model with Word2Vec to create our vectors to be ready feed in to our machine learning algorithms.

#### Initial K-means to compute optimal clusters

K-means clustering is well known machine learning algorithm that’s prominently used on unsupervised data like text and image content. Here the algorithm works on the unknown data which does have label or outcomes. For example the vectors that represent words after word embeddings does not have any label and is perfect algorithm to make use in our proposed approach. With literature review on various we came to know that Chadha, A [Chadha, A,2014] K-means has two major set back when we apply on unsupervised data that is compute initial centroids of cluster and selection of optimal(k) number of clusters to be formed.

* We are proposing below techniques to mitigate this set back of K-means algorithm by tuning parameters of K-Means will start cluster Range from 4 to N for better cluster density to overcome initial sparse cluster centroid computation. This will help us to have better cluster centers.
* For selection of optimal k-value we have proposed silhouette Method to determine the k value.

Silhouette Method uses silhouette value measures  - means how similar a point is to its own cluster (cohesion) compared to other clusters (separation).The range of the Silhouette value is between +1 and -1. A high value is desirable and indicates that the point is placed in the correct cluster. If many points have a negative Silhouette value, it may indicate that we have created too many or too few clusters. The Silhouette Value ***s(i)*** for each data point ***i*** is defined as follows:



Source: Wikipedia

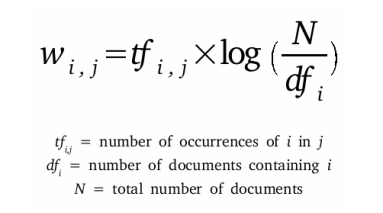
Hence based the average Silhouette score of each cluster we will determine best cluster and select best cluster to determine value of optimal(k)number of clusters.

#### Latent Semantic Analysis using TF-IDF

Latent semantic analysis (LSA) is way to obtain distributional semantics, by analysing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms.

Term Frequency — Inverse Document Frequency (TF-IDF) is one of the underlying technique in LSA that will be used to create the word vector representations.

Given m documents and n-words in our text content, we can create an  
m × n matrix A in which each row represents a document and each column represents a word



Picture-1: Term Frequency- Inverse Document Frequency (Tf-IDF) Score

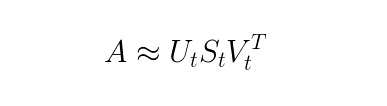
Hence a document-term matrix, A was formed using this transformation method(tf-IDF) to vectorize our text content.

#### Apply K-means algorithm on Truncated SVD

In this step will be forming clusters by using K-Means clustering algorithm based on the optimal(k) number of cluster that was computed earlier in 3.3.3.3 section.

In tf-IDF vector the term which occur more will have more weight that the less frequently occurring words, so it is prone to form noisy and sparse matrix.

In order to mitigate this noise will can reduce the dimensionality of the matrix by using **Truncated Singular Value Decomposition (SVD).** **This is a *Low-Rank Approximation technique which*  selects only the *t* largest singular values.** Singular value decomposition is a technique in linear algebra that factorizes any matrix M into the product of 3 separate matrices: M=U\*S\*V, where S is a diagonal matrix of the singular values of M.



**Hence based on the text content we can choose t value of the truncated SVD to reduce the dimensionality of the vector matrix.**

**Will use this Truncated SVD Tf-IDF as input to our K-means algorithm with computed K-value and generate clusters and then save to file/disk for further processing.**

#### Labelling Cluster Content using LDA

#### LDA hyperparameter Tuning

### Topic visualization within cluster

In this Initial K-Means Algorithm we will form initial clusters of the text data using K-means clustering technique with specific cluster range instead of starting with 2 to N ,we have chosen to start from 4 to N to improve initial parse clusters and form better density with in data points.

While forming these cluster we will calculate the silhouette scores of each cluster for training model and then extract best silhouette score clusters to form final cluster clouds and split that data into respective unlabelled cluster corpus files.

Below are detailed steps involved in this algorithm –

**Input**: Employee-review data in csv format will be the text corpus we will be forming clusters

**Output**: Distinctive suitable number of cluster corpuses based on number of best clusters formed as a result

**Method**:

1. Collecting text data from external source and import your runtime environment for further processing
2. We use various techniques to clean the data, depending on the domain of the problem. Regular expressions are extensively used at this stage in identifying anomalies in the data and transforming them as required
3. Normalization of text data includes techniques such as "removal of stop words", "converting all text to a standard case", "stemming/lemmatization”
4. Feature engineering textual data typically involves converting the cleaned and normalized data into vector representations. These vectors can then be used by machine learning algorithms to build models, we have chosen K-means based on our literature review for forming initial clusters
5. Instead of starting cluster range from 2 to N will be skipping beginning clusters and start converging our data from 4 to N so that better cluster are formed instead of sparse cluster which are sensitive to outliers.
6. Will be calculating the silhouette scores during step 5 to chose best clusters with better silhouette score. These best cluster give us the optimal numbers (K) of cluster to be created. Silhouette score here refers to a method of interpretation and validation of consistency within clusters of data. The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters.
7. Then we will fit that best clusters selected to form the distinctive cluster text corpus which are unlabelled cluster.

Refer Figure 1.0 for Design Flow chart of Proposed Approach-I and Proposed Approach-II.

Then we will make use of LDA topic modelling technique on these extracted cluster corpuses. By doing hyperparameter tuning for LDA parameters will calculate the optimal number of topics with each cluster corpus by selecting optimal alpha and beta parameters to form better coherence measures.

Hence by this we can achieve our objective of extracting the key factors hidden in employee text corpus.

3.2 Labelling Cluster Content using LDA

Label clustering is also can be referred as topic modelling to name the features within the cluster corpuses which was generated in previous section.

We will be using Latent Dirichlet Allocation (LDA) topic modelling for finding the patterns within each cluster corpuses.

Figure 2.0 for Flow chart for Label clustering using LDA

Below are the detailed steps of the Labelling Cluster Content using LDA:

1.) Retrieve respective cluster file for processing.

2.) Data Cleaning and normalization of text data.

3.) Pre-processing Using POS tagging using.

4.) Use Basic LDA method with random values for number of topics with each cluster corpus by selecting optimal alpha and beta.

5.) Apply hyper tuning techniques to determine optimal number of topics with each cluster.

6.) Visualize the final features in each cluster corpus.

Then we can label the features extracted from the topic with estimated term frequency within the selected clusters.

## 3.4 **Summary**

As the size of the corpus grows with each cluster the estimated term frequency within the selected cluster might be sensitive with lesser coherence scores. Hence, we need apply better hyper parameters to improve the coherence score by choosing optimal topic which are distinct.

# Chapter 4

**ANALYSIS**

## 4.1 Introduction

## 4.2 Dataset Description

The brief description of data and its inter-relation with other objects within data set is called Data Dictionary/Dataset Description.

This data set was pulled collected from glass-door employee review site by Kaggle and is public domain data with any individual or organization information.

Below are the columns of the dataset that we have chosen.

* 'ID': Identification Number
* 'Place': Startup {1-6}
* 'location': Location of startup
* 'date': Date of review
* 'status': Current status with the startup
* 'job\_title': Position of work at the startup
* 'summary': Overall summary
* 'positives': Pros
* 'negatives': Cons
* 'advice\_to\_mgmt': Comments given by the reviewer to the management
* 'overall': Overall rating provided by the user {1-5}
* 'score\_1' to 'score\_5': Intricate rating with reflects the condition of work at the startup {1-5}
* 'score\_6': Number of likes received by the reviewer for the review.

## 4.3 Data Set Preparation

## 4.4 Exploratory Data Analysis

## 4.5 Data Visualization

## 4.5 Summary

4.1 Dataset

**4.1.1 Data Set Preparation**

Data source from glass-door is referred from Kaggle which has raw test dataset of actual employee review texts at various organizations in market. This data has many details like ID, startup location, job title, summary of review, pros, cons, advice to management and overall rating. As part of this research we primarily focus on the pros, cons and advice to management columns which has unstructured text reviews with train and test data segregation organized. We have considerate 30336 rows of reviews collected for in train set and 29272 rows in test set.

**4.1.2 Data Set Analysis**

*Columns*

'ID': Identification Number 'Place': Start-up {1-6} 'location': Location of startup 'date': Date of review 'status': Current status with the startup 'job\_title': Position of work at the startup 'summary': Overall summary 'positives': Pros 'negatives': Cons 'advice\_to\_mgmt': Comments given by the reviewer to the management 'overall': Overall rating provided by the user {1-5} 'score\_1' to 'score\_5': Intricate rating with reflects the condition of work at the startup {1-5} 'score\_6': Number of likes received by the reviewer for the review

Data quality is risk as motivation of this research is based on qualitative data from the survey conducted with small sample of employees as part of primary data source but here due to time constraints am using dataset from public domain which might provide data with difference experience level with varied reviews.

4.2 Results Analysis

***Source code link is available in appendix section and initial Experiment results are available in figure 3.1 and 3.2.***

---------Work in progress------

4.3 Error Analysis

---------Work in progress------

# Chapter 5

**RESULTS AND DISCUSSIONS**

## 5.1 Introduction

## 5.2 Interpretation of Visualizations

## 5.3 Evaluation of Results

## 5.4 Summary

# Chapter 6

**CONCLUSION & FUTURE WORK**

## 6.1 Introduction

## 6.2 Discussion and Conclusion

In thesis a combination of Initial K-means algorithm with LDA is proposed as optimized way of choosing optimal cluster number based on best silhouette scores. Then with the help of topic modelling using Latent Dirichlet Allocation (LDA) method. So far, the experimental results are better than the traditional K-means algorithm. Also, the computation time of the proposed algorithm Model-II using Truncated SVD is better than the Model-I using Word2Vec embedding.

I am in process of LDA hyper tuning and results analysis which will be continued in the next few days.

## 6.3 Contribution to knowledge

## 6.4 Future Recommendations

Future work

We would like to build a comprehensive model that can be used on real-time to retrieve document cluster lookup for the key features within employee review text.

# List of figures

Figure 1.0 for Design Flow chart of Proposed Approach-I and Proposed Approach-II

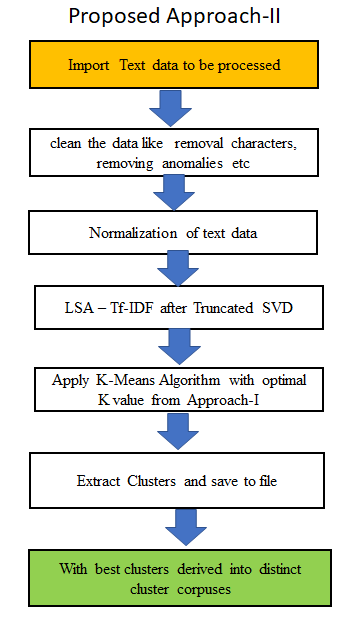
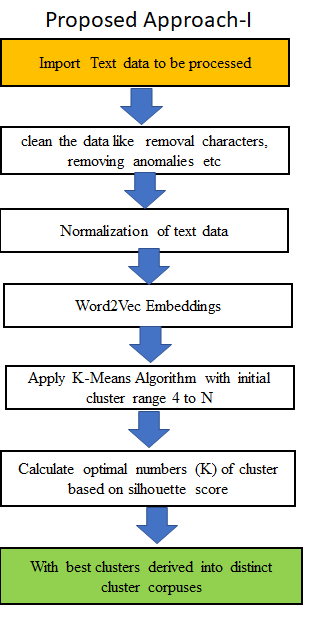


Figure 2.0 for Flow chart for Labelling Cluster Content using LDA

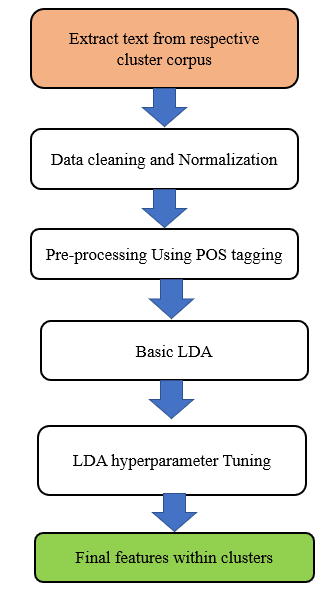
****

Figure 3.0 for Applying Initial K-Means Clustering output

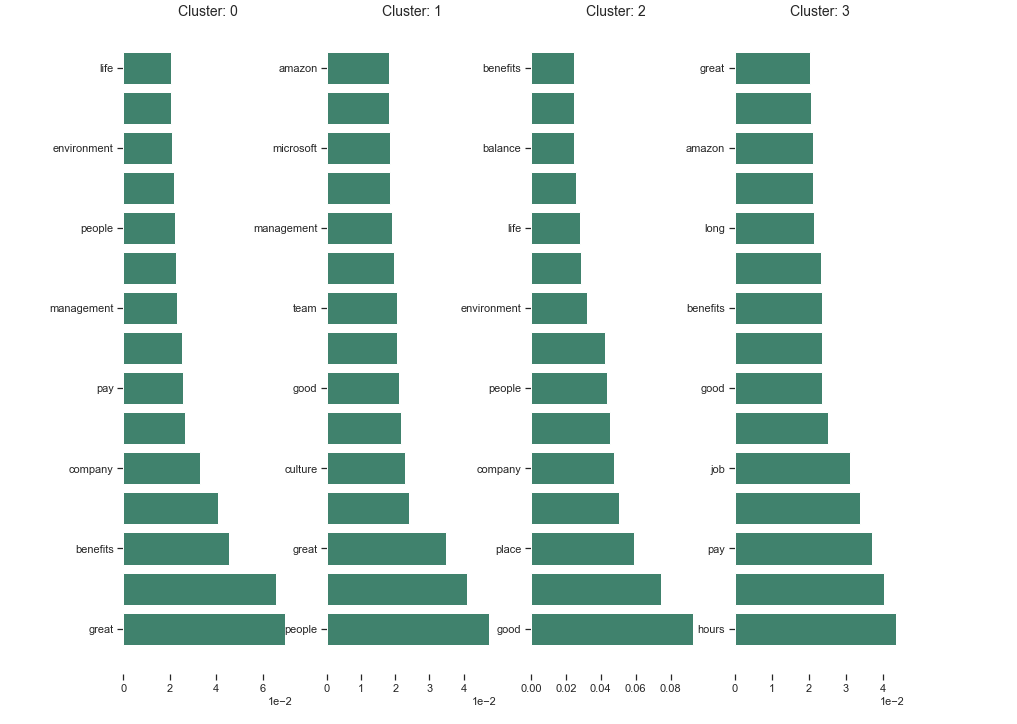


Figure 4.0 for Topic modelling using LDA

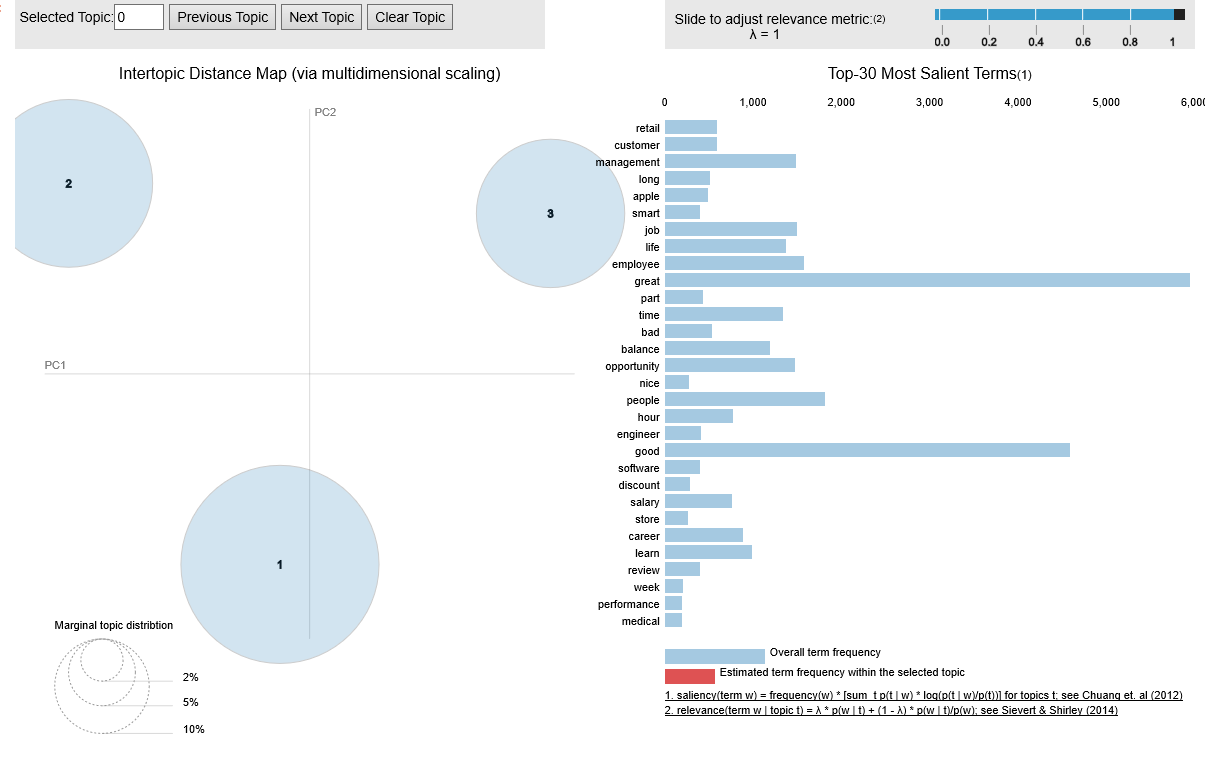
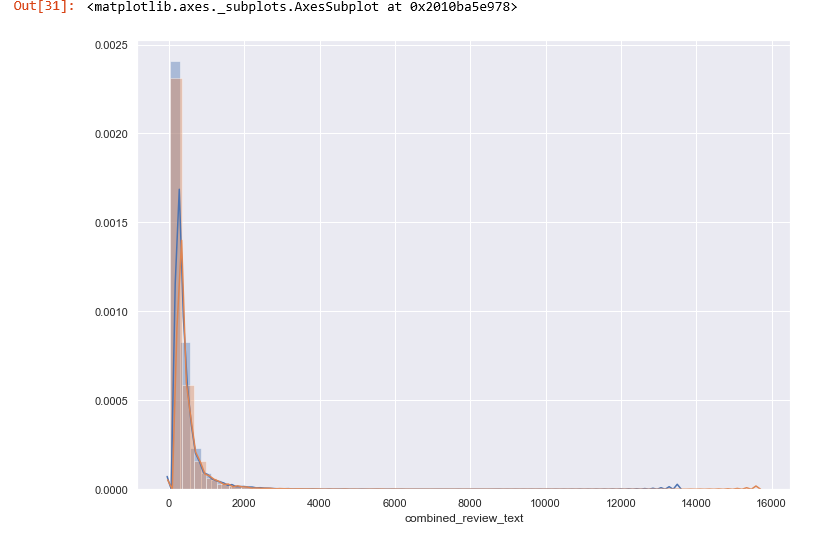


Figure 5.0 Data Visualization

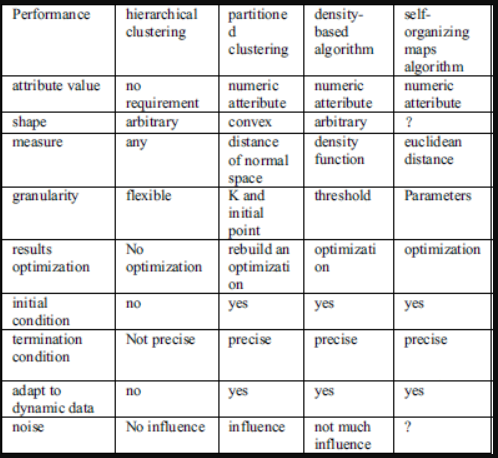


Figure 6.0 Word Frequency Distribution in Document

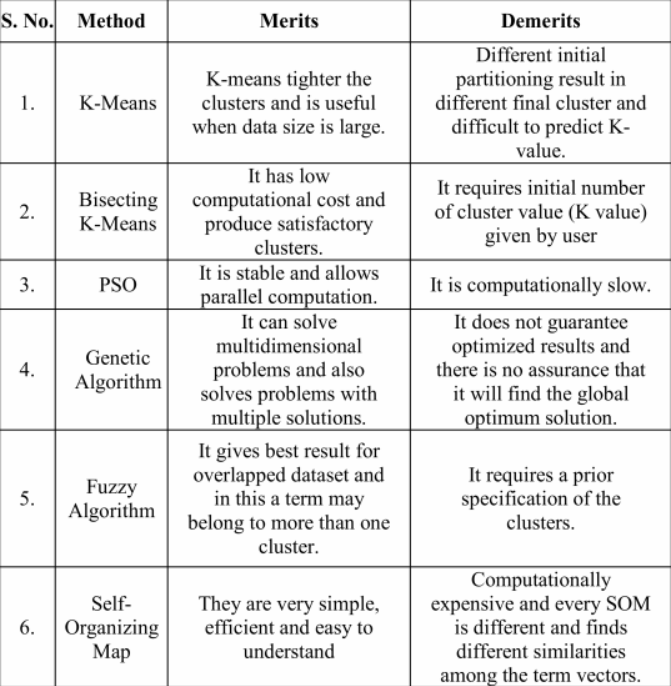


# list of tables

**Table 1.0**Performance Comparison of Clustering Algorithms



**Table 2.0**Merits and demerits of clustering algorithms



# REFERENCES

Aggarwal, C. C., & Zhai, C. X. (2012). A survey of text clustering algorithms. In *Mining Text Data*. https://doi.org/10.1007/978-1-4614-3223-4\_4

Ahmad, A., Ruhul Amin, M., & Chowdhury, F. (2018). Bengali Document Clustering Using Word Movers Distance. *2018 International Conference on Bangla Speech and Language Processing, ICBSLP 2018*. https://doi.org/10.1109/ICBSLP.2018.8554598

Bancken, W., Alfarone, D., & Davis, J. (2014). Automatically detecting and rating product aspects from textual customer reviews. *CEUR Workshop Proceedings*.

Berry, M. W., & Castellanos, M. (2008). Survey of text mining II: Clustering, classification, and retrieval. In *Survey of Text Mining II: Clustering, Classification, and Retrieval*. https://doi.org/10.1007/978-1-84800-046-9

Bhutada, S., Balaram, V. V. S. S. S., & Bulusu, V. V. (2014). Latent Dirichlet Allocation based multilevel classification. *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies, ICCICCT 2014*. https://doi.org/10.1109/ICCICCT.2014.6993109

Bijalwan, V., Kumar, V., Kumari, P., & Pascual, J. (2014). KNN based machine learning approach for text and document mining. *International Journal of Database Theory and Application*. https://doi.org/10.14257/ijdta.2014.7.1.06

Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. a., & Reynar, J. (2008). Building a Sentiment Summarizer for Local Service Reviews. *WWW Workshop on NLP in the Information Explosion Era*. https://doi.org/10.1109/MIS.2013.30

Bouraoui, A., Jamoussi, S., & Hamadou, A. Ben. (2018). A new method for the construction of evolving embedded representations of words. *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*. https://doi.org/10.1109/AICCSA.2017.155

Boutsidis, C., & Magdon-Ismail, M. (2013). Deterministic feature selection for k-means clustering. *IEEE Transactions on Information Theory*. https://doi.org/10.1109/TIT.2013.2255021

Chadha, A., & Kumar, S. (2014). An improved K-Means clustering algorithm: A step forward for removal of dependency on K. *ICROIT 2014 - Proceedings of the 2014 International Conference on Reliability, Optimization and Information Technology*. https://doi.org/10.1109/ICROIT.2014.6798312

Charu C. Aggarwal, C. K. R. (2013). *DATA CLUSTERING Algorithms and Applications*.

Devika, M. D., Sunitha, C., & Ganesh, A. (2016). Sentiment Analysis: A Comparative Study on Different Approaches. *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2016.05.124

Fasheng Liu, & Lu Xiong. (2011). *Survey on text clustering algorithm*. https://doi.org/10.1109/icsess.2011.5982485

Ferreira, R., Lins, R. D., Freitas, F., Avila, B., Simske, S. J., & Riss, M. (2014). A new sentence similarity method based on a three-layer sentence representation. *Proceedings - 2014 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT 2014*. https://doi.org/10.1109/WI-IAT.2014.23

Gupta, A., Gautam, J., & Kumar, A. (2018). A survey on methodologies used for semantic document clustering. *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing, ICECDS 2017*. https://doi.org/10.1109/ICECDS.2017.8389521

Güven, Z. A., Diri, B., & Çakaloǧlu, T. (2018). Classification of New Titles by Two Stage Latent Dirichlet Allocation. *Proceedings - 2018 Innovations in Intelligent Systems and Applications Conference, ASYU 2018*. https://doi.org/10.1109/ASYU.2018.8554027

Hsu, C. I., & Chiu, C. (2017). A hybrid Latent Dirichlet Allocation approach for topic classification. *Proceedings - 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications, INISTA 2017*. https://doi.org/10.1109/INISTA.2017.8001177

Hu, K., Wu, H., Qi, K., Yu, J., Yang, S., Yu, T., … Liu, B. (2018). A domain keyword analysis approach extending Term Frequency-Keyword Active Index with Google Word2Vec model. *Scientometrics*. https://doi.org/10.1007/s11192-017-2574-9

Huang, X., Peng, F., Schuurmans, D., Cercone, N., & Robertson, S. E. (2003). Applying machine learning to text segmentation for information retrieval. *Information Retrieval*. https://doi.org/10.1023/A:1026028229881

Jawahar, G., Sagot, B., & Seddah, D. (2019, September 12). *What Does BERT Learn about the Structure of Language?* 3651–3657. https://doi.org/10.18653/v1/p19-1356

Jianqiang, Z., & Xiaolin, G. (2017). Comparison research on text pre-processing methods on twitter sentiment analysis. *IEEE Access*. https://doi.org/10.1109/ACCESS.2017.2672677

Karl, A., Wisnowski, J., & Rushing, W. H. (2015). A practical guide to text mining with topic extraction. *Wiley Interdisciplinary Reviews: Computational Statistics*. https://doi.org/10.1002/wics.1361

Khan, S. S., & Ahmad, A. (2013). Cluster center initialization algorithm for K-modes clustering. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2013.07.002

Khan, S. S., & Ahmad, A. (2017). Cluster center initialization algorithm for K-means clustering Cluster center initialization algorithm for K -means clustering. *Pattern Recognition Letters*. https://doi.org/10.1016/j.patrec.2004.04.007

Kim, S., Zhang, J., Chen, Z., Oh, A., & Liu, S. (2013). A hierarchical aspect-sentiment model for online reviews. *Proceedings of the 27th AAAI Conference on Artificial Intelligence, AAAI 2013*.

Li, X. (2008). Research on text clustering algorithm based on K-means and SOM. *Proceedings - 2nd 2008 International Symposium on Intelligent Information Technology Application Workshop, IITA 2008 Workshop*. https://doi.org/10.1109/IITA.Workshops.2008.13

Li, Y., Algarni, A., Albathan, M., Shen, Y., & Bijaksana, M. A. (2015). Relevance feature discovery for text mining. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2014.2373357

Marrese-Taylor, E., Velásquez, J. D., Bravo-Marquez, F., & Matsuo, Y. (2013). Identifying customer preferences about tourism products using an aspect-based opinion mining approach. *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2013.09.094

Pardeshi, B., & Toshniwal, D. (2010). Improved K-medoids clustering based on cluster validity index and object density. *2010 IEEE 2nd International Advance Computing Conference, IACC 2010*. https://doi.org/10.1109/IADCC.2010.5422924

Park, H. S., Lee, J. S., & Jun, C. H. (2006). A k-means-like algorithm for k-medoids clustering and its performance. *36th International Conference on Computers and Industrial Engineering, ICC and IE 2006*.

Peng, D., Wu, S., & Liu, C. (2019). MPSC: A Multiple-Perspective Semantics-Crossover Model for Matching Sentences. *IEEE Access*. https://doi.org/10.1109/ACCESS.2019.2915937

Reddy, D., & Jana, P. K. (2012). Initialization for K-means Clustering using Voronoi Diagram. *Procedia Technology*. https://doi.org/10.1016/j.protcy.2012.05.061

Singh, T., & Kumari, M. (2016). Role of Text Pre-processing in Twitter Sentiment Analysis. *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2016.06.095

Srikanth, P., & Deverapalli, D. (2018). CFTDISM:Clustering Financial Text Documents Using Improved Similarity Measure. *2017 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2017*. https://doi.org/10.1109/ICCIC.2017.8524466

Steven Bird, Ewan Klein, and E. L. (2009). *Natural Language Processing with Python*. O’Reilly Media.

Su, T., & Dy, J. (2004). A deterministic method for initializing K-means clustering. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*. https://doi.org/10.1109/ICTAI.2004.7

Tunali, V., & Bilgin, T. T. (2012). Examining the impact of stemming on clustering Turkish texts. *INISTA 2012 - International Symposium on INnovations in Intelligent SysTems and Applications*. https://doi.org/10.1109/INISTA.2012.6246966

Twinandilla, S., Adhy, S., Surarso, B., & Kusumaningrum, R. (2018). Multi-Document Summarization Using K-Means and Latent Dirichlet Allocation (LDA) - Significance Sentences. *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2018.08.220

Usop, E. S., Isnanto, R. R., & Kusumaningrum, R. (2018). Part of speech features for sentiment classification based on Latent Dirichlet Allocation. *Proceedings - 2017 4th International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2017*. https://doi.org/10.1109/ICITACEE.2017.8257670

Xia, R., Xu, F., Zong, C., Li, Q., Qi, Y., & Li, T. (2015). Dual Sentiment Analysis: Considering Two Sides of One Review. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2015.2407371

Xinwu, L. (2012). A new text clustering algorithm based on improved k\_means. *Journal of Software*. https://doi.org/10.4304/jsw.7.1.95-101

Xiong, C., Hua, Z., Lv, K., & Li, X. (2017). An improved K-means text clustering algorithm by optimizing initial cluster centers. *Proceedings - 2016 7th International Conference on Cloud Computing and Big Data, CCBD 2016*. https://doi.org/10.1109/CCBD.2016.059

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*. https://doi.org/10.1109/TNN.2005.845141

Xu, X., & Ye, F. (2018). Sentences similarity analysis based on word embedding and syntax analysis. *International Conference on Communication Technology Proceedings, ICCT*. https://doi.org/10.1109/ICCT.2017.8359959

Yu, D., Liu, G., Guo, M., & Liu, X. (2018). An improved K-medoids algorithm based on step increasing and optimizing medoids. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2017.09.052

Yu, S. S., Chu, S. W., Wang, C. M., Chan, Y. K., & Chang, T. C. (2018). Two improved k-means algorithms. *Applied Soft Computing Journal*. https://doi.org/10.1016/j.asoc.2017.08.032

Yuan, S., & Guo, W. (2013). A text clustering algorithm based on simplified cluster hypothesis. *Proceedings - 2013 2nd International Symposium on Instrumentation and Measurement, Sensor Network and Automation, IMSNA 2013*. https://doi.org/10.1109/IMSNA.2013.6743303

Zamir, O. E. (1999). *Clustering Web Documents : A Phrase-Based Method for Grouping Search Engine Results*.

Zhang, X., & Yu, Q. (2017). Hotel reviews sentiment analysis based on word vector clustering. *2017 2nd IEEE International Conference on Computational Intelligence and Applications, ICCIA 2017*. https://doi.org/10.1109/CIAPP.2017.8167219

Zhenzhong Lan and Mingda Chen and Sebastian Goodman and Kevin Gimpel and Piyush Sharma and Radu Soricut.(2019).ALBERT: A Lite BERT for Self-supervised Learning of Language Representations.arXiv:1909.11942 [cs.CL]

Zhou, L., Wang, L., Ge, X., & Shi, Q. (2010). A clustering-based KNN improved algorithm CLKNN for text classification. *CAR 2010 - 2010 2nd International Asia Conference on Informatics in Control, Automation and Robotics*. https://doi.org/10.1109/CAR.2010.5456668

Zhou, Y., Li, Y., & Xia, S. (2009). An improved KNN text classification algorithm based on clustering. *Journal of Computers*. https://doi.org/10.4304/jcp.4.3.230-237

Zhu, M., Wang, W., & Huang, J. (2014). Improved initial cluster center selection in K-means clustering. *Engineering Computations (Swansea, Wales)*. https://doi.org/10.1108/EC-11-2012-0288

Zhuang, Y., Gao, H., Wu, F., Tang, S., Zhang, Y., & Zhang, Z. (2015). Probabilistic word selection via topic modeling. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2014.2377727

**Additional References:**

[OR.1] [https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/#](https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/)

[OR.2] <https://www.slideshare.net/Micro-Focus/growth-of-internet-data-2017>

[OR.3] <https://octoverse.github.com/>

[OR.4] <https://www.analyticsvidhya.com/blog/2019/02/flair-nlp-library-python/>

[OR.5] https://blog.algorithmia.com/introduction-natural-language-processing-nlp/

[OR.6] https://towardsdatascience.com/introduction-to-natural-language-processing-for-text-df845750fb63

[OR.7]<https://medium.com/activewizards-machine-learning-company/comparison-of-top-6-python-nlp-libraries-c4ce160237eb>

[OR.8]<https://www.quora.com/What-are-the-main-differences-between-the-word-embeddings-of-ELMo-BERT-Word2vec-and-GloVe>

# APPENDIX A: RESEARCH PROPOSAL



# APPENDIX B: ETHICS FORMS

# APPENDIX C: Additional References

1. Dataset – <https://www.kaggle.com/fireball684/hackerearthericsson>
2. The source code can be found in below GitHub link for reference –

<https://github.com/goudkedar/NLP_thesis_code>