2020

A Comparative Study of Text Summarization on E-mail Data Using Unsupervised Learning Approaches

Tijo Thomas

Technological University Dublin

A comparative study of Text summarization on E-mail data using Unsupervised learning approaches

Tijo Thomas

Technological University Dublin

A dissertation submitted in partial fulfilment of the requirements of
Technological University Dublin for the degree of
M.Sc. in Computing (Data Analytics)

2019

DECLARATION

I, Tijo Thomas certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Technological University of Dublin and has not been submitted in whole or part for an award in any other Institute or University.

The work reported in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: Tijo Thomas

Date: 01 January 2020

ABSTRACT

Over the last few years, email has met with enormous popularity. People send and receive a lot of messages every day, connect with colleagues and friends, share files and information. Unfortunately, the email overload outbreak has developed into a personal trouble for users as well as a financial concerns for businesses. Accessing an ever-increasing number of lengthy emails in the present generation has become a major concern for many users. Email text summarization is a promising approach to resolve this challenge. Email messages are general domain text, unstructured and not always well developed syntactically. Such elements introduce challenges for study in text processing, especially for the task of summarization.

This research employs a quantitative and inductive methodologies to implement the Unsupervised learning models that addresses summarization task problem, to efficiently generate more precise summaries and to determine which approach of implementing Unsupervised clustering models outperform the best. The precision score from ROUGE-N metrics is used as the evaluation metrics in this research. This research evaluates the performance in terms of the precision score of four different approaches of text summarization by using various combinations of feature embedding technique like Word2Vec /BERT model and hybrid/conventional clustering algorithms. The results reveals that both the approaches of using Word2Vec and BERT feature embedding along with hybrid PHA-ClusteringGain k-Means algorithm achieved increase in the precision when compared with the conventional k-means clustering model. Among those hybrid approaches performed, the one using Word2Vec as feature embedding method attained 55.73% as maximum precision value.

Keywords: Text Summarization, Electronic mail, Unsupervised Learning, PHA-Clustering Gain, K-means clustering, ROUGE

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my Supervisor **Brian Leahy**, for providing invaluable guidance and tremendous support throughout the dissertation process. It was a great honor to work and study under his supervision.

I submit my heartiest gratitude to **Dr. Luca Longo** for the guidance and insightful comments. Without his diligent help, the research for this dissertation would not have been complete.

On this moment I remember all my teachers from childhood who have always been my motivation in learning and for transferring the knowledge as well as guiding me in this journey of life.

I also want to thank my beloved parents for being my inspiration. Special thanks to all the people who along the way believed in me.

My special appreciation to my classmates and friends in providing relevant assistance and help to complete this study.

Contents

List of Figures

List of Tables

List of Acronyms

ML	Machine Learning
NLP	Natural Language Processing
CRISP-DM	Cross Industry Standard Process for Data Mining
BERT	Bidirectional Encoder Representations from Transformers
MLM	Masked Language Model
NSP	Next Sentence Prediction
PHA	Potential-based Hierarchical Agglomerative Clustering
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
LSI	Latent Semantic Indexing
PLSA	Probabilistic Latent Semantic Analysis
NMF	Non-negative Matrix Factorization
HMM	Hidden Markov Model
HAC	Hierarchical Agglomerative Clustering
NER	Named Entity Recognition
CBOW	Continuous Bag of Words

Chapter 1

INTRODUCTION

1.1 Background

Radev et al. (2002) define a summary as "a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that". Text Summarization is process of shortening source text into a small version by preserving the content of information and the meaning of the context. The surplus availability of information in unstructured format have intensively necessitated the research in the area of automatic text summarization. Due to the inability of people to assimilate vast amounts of information, efficient methods of summarizing text are important with the explosion of data available on the Web and other media in the form of unstructured text. Most of the text summarization methods typically engages with multiple approaches either to identify the most relevant sentences in the text or to remove those sentences which are redundant and irrelevant. Automatic Text Summarization shortens the volume of information by creating a summary from one or more paragraph of sentences or from text documents without losing any of the main contents in it. The focus of the summary may be either generic, which captures the important semantic concepts of the documents, or query-based, which captures the sub-concepts of the user query and provides personalized and relevant abstracts based on the matching between input query and document collection. Because of the enormous load of information generated especially on the web, including large text, audio, and video files, etc., summarization has gained interest

in research. Text Summarization develops an overview of a large texts/documents which allows the user to understand and reject/include the text without reading the whole text. Summarization can also be useful in classifying text, answering questions, retrieving data, etc. These summarization techniques are widely used in search engines like google in order to improve the quality of search.

Summarization techniques are broadly classified into extractive and abstractive summarization. Extractive summarization methods identify relevant sentences from the original text and string them together to form a summary. Abstractive summarization methods are those that can generate summary sentences that are not present in the original text [Das, D., Martins, A., (2007)]. There are several extractive text summarization techniques which have been used from the late fifties till now and those can be broadly classified based on its nature into five methods: Statistical, Fuzzy-Logic, Graph, Machine-learning and Latent Semantics approach and additionally into topic, discourse method that are based on one or more from the previous approaches. These approaches can be classified depending on different learning type into supervised, semi-supervised and unsupervised approach. The below Figure 1. displays the extractive text summarization techniques categorized based on learning type.

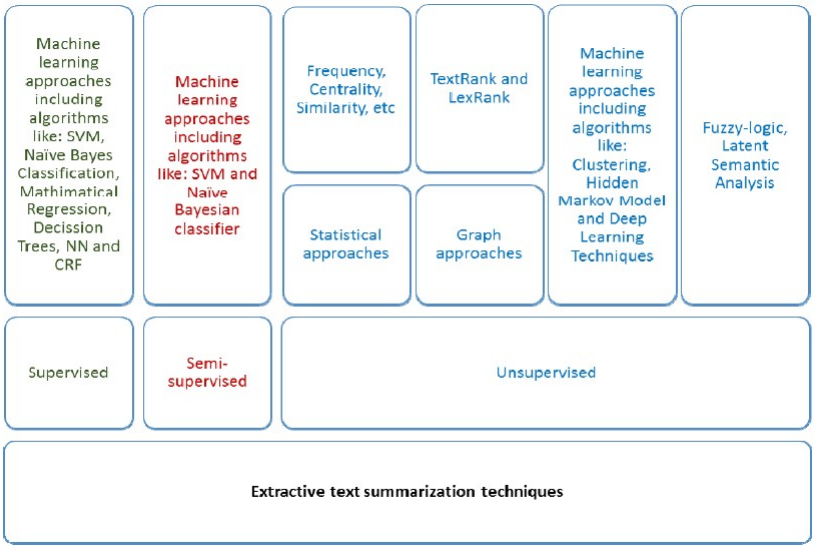


Figure 1.1: Classification of Extractive Text Summarization Techniques

I propose an approach to perform an extractive text summarization using Unsupervised Learning algorithms and compare multiple variants of clustering based on the performance.

This can be hypothetically structured as :

H0 :The application of PHA-ClusteringGain k-Means hybrid approach in text summarization will result in no precision increase over a conventional k-means clustering model.

H1 :The application of PHA-ClusteringGain k-Means hybrid approach will lead to an increase of precision rate in text summarization over a conventional k-means clustering model.

1.2 Research problem

The impact of electronic mail is now more evident than ever in our daily lives. Millions of plain text or enriched messages are sent and received around the globe every minute. Some of them are read with extra care and at the same time, with obvious disinterest, many of them are deleted. As the internet grows, electronic mail has become not only a vital tool for our work, but also an important means of communication between people. Email also significantly facilitated personal communication as it offered instant messaging with minimal cost. People from around the world can now exchange views and information so easily that email has become the second most popular channel of voice after voice communications. Features that made email so popular are the speed of communication, the minimum cost, and the remarkably easy use of it. An advantage over voice communication (e.g. telephone) is that it is asynchronous, meaning that both sides of communication do not need to be simultaneously online or in front of a computer. Unfortunately, the curse of Information Overload could not escape email. Loads and loads of incoming messages have turned the handling of electronic mail into a tedious task (some extremely important, others simply junk). So, with the application of Text summarization on electronic mail, we can extensively save the time as well as retrieve the precised meaning.

The aim of this research is to implement the text summarization using various unsupervised learning techniques like a hybrid approach of PHA-ClusteringGain-KMeans algorithm and conventional K-means algorithm on Email data by applying different feature embedding techniques like word2vec and BERT model. The objective of the research is to

evaluate the performance interms of precision of the hybrid clustering model compared to the conventional algorithm with different approaches. Since this research centers on a clustering problem, the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) method is used for evaluating automatic summarization which helps in accessing the performance and the precision of the models developed in the research.
To guide the research, the research question has been structured as:

"To what extent can a PHA-Clustering Gain k-Means hybrid approach out-perform a conventional k -means clustering model for email-based text summarization ?"

1.3 Research Objectives

The key objective of this research is to determine whether the application of hybrid approach from unsupervised machine learning on text summarization will out-perform a conventional unsupervised learning model in terms of determining the performance and comparing the precision.

Hypothetically it can be outlined as : "If the precision rate of PHA-ClusteringGain-KMeans hybrid model is found less than conventional k -means clustering model , then we accept the null hypothesis (H0) and reject the alternate hypothesis (H1). While if PHA-ClusteringGain-KMeans hybrid approach out-perform the conventional k -means clustering model, then we accept the alternate hypothesis (H1) and reject the null hypothesis (H0)". The principal objective is to plan and execute experiments that seek to reject the null hypothesis.

The objectives associated with this research are:

1. To review related literature on text summarization, Unsupervised machine learning models, tokenization, sentence encoding and summary evaluation.

2. To gather the necessary data which is required for the research.

3. Identify and fix any errors or issues in terms of the quality of the data, that can affect performance of the machine learning models.

4. Prepare the data by data cleaning, encoding and feature extraction.

5. Apply different embedding techniques like word2vec and BERT on the data before processing.

6. Build the machine learning models using unsupervised learning algorithms such as PHA-ClusteringGain-KMeans hybrid model and conventional K- means model.

7. To perform a comparative study of the models based on its performance and precision rate.

8. Evaluate the performance of the models using ROUGE method as the evaluation metrics.

9. Recognize the limitations of this research and suggest areas of future research.

1.4 Research Methodologies

A Quantitative research methodology by Objective is used in this research because it involves conducting experiments on the email data to build machine learning models and measure the precision and the outperformance between the algorithms. The hypothesis is accepted/rejected based on the summarized precision of different machine learning algorithms and the result of the research is based on the experiments and a comparison is made between the precision of the machine learning models, making the reasoning of the research as Inductive. Since this research is collation to an existing research, making this as a Secondary research by Type. As part of experimental study, the data was split into a training and a test set and a comparative study between a hybrid machine learning algorithm and a conventional algorithm for this research problem was carried out using these sets. In this research another point was to gain knowledge by comparison of performance of algorithms with hypotheses testing and suitable experiments, making the research as Empirical by form.

Cross Industry Standard Process for Data Mining (CRISP - DM) methodology is used in this research. It provides a structured method to execute a data mining project. It is a robust

and well-proven methodology that consists of six phases, which are business understanding, data understanding, data preparation, modelling, evaluation and deployment.

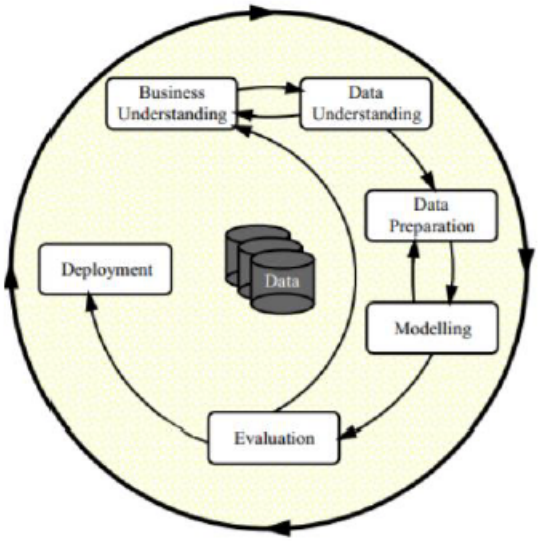


Figure 1.2: CRISP DM Lifecycle (Source: Wirth Hipp, 2000)

1.5 Scope and Limitations

There are many aspects and research regarding automatic text summarization that, apart from their importance, cannot be investigated. This research work is to a wider scope to develop text summarization using different unsupervised learning methods. The scope of this study is to develop different machine learning models using the email data to generate the summarized contents form the email dataset and to perform a comparative study of both the models based on its performance and precision rate.

Chaining is a common problem in Single Linkage clustering and PHA method and it can be defined as the gradual growth of a single cluster as one data object with the elements added to that cluster at each iterative step of the algorithm. This leads to the formation of impractical heterogeneous clusters and may result in the unequal partitioning of data objects. In a clustering process many singleton clusters are formed as a result of chaining. Thus, the output clusters cannot be properly defined from the input data objects.

The major limitation of the research is deciding optimum number of clusters k and it is

the major problem with Partitioning-based clustering approaches such as k-means method. But the proposed hybrid approach helps to overcome this limitation. Another limitation to get added up is regarding the data. There can be some email records with limited number of words or sentences, hence as part of implementing text summarization we have used only those emails records with more than 329 words(mean wordcount) which stays as a threshold for the minimum number of words being used for the research.

Another main limitation is rather challenging to get a public email data for researches and studies, due to privacy related issues and GDPR establishment. This is a drawback and a limitation especially for research since the studies cannot be conducted due to unavailability of public email datasets. An exemption to the beyond challenge is the Enron Corpus [Klimt, Yang, (2004)], in which this email data was made public after a legal investigation regarding the Enron Corporation.

1.6 Document Outline

The rest of the thesis is organized as follows:

Chapter 2 (Literature review and related work) discuss the literature related to the various Unsupervised learning methods, Chaining effect, Sampling methods, Clustering algorithms especially PHA-ClusteringGain-KMeans hybrid model and conventional k -means clustering model.

Chapter 3 (Design and Methodology) discuss the design of the research in deep and how each phase of CRISP- DM methodology is followed in the research is discussed in detail here. In addition to that, the process of cleaning the data, engineering the data, training the machine learning models using the data, evaluation of the models depending on the performance and precision of models and implementation of the best model are discussed in this chapter.

Chapter 4 (Implementation and Results) outlines the complete Implementation of all the experimental approaches and presents the result of the respective experiments as per the proposed design.

Chapter 5 (Evaluation and discussion) perform the evaluations of the experiments in terms of class precision and a comparative study of the results from each of the experiments and its discussion.

Chapter 6 (Conclusion) summarizes the research that is carried out and discusses how the experiments and the following results answers the research question which was discussed in the Chapter 1. The chapter concludes with discussing areas of future research.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

Chapter 2 reviews the research literature which was done in the field of knowledge discovery, data mining and machine learning in particular clustering a form of unsupervised learning and different algorithms used for clustering. The purpose of this research and the experimental approach outlined below in Chapter 3 is to extend the existing part of research in this section.

In Section 2.2, The state of the art frameworks/methodologies for knowledge discovery and machine learning are reviewed based on three main forms of learning techniques, supervised, unsupervised and reinforcement learning.

In section 2.3, A brief discussion on implementation and growth of machine learning and technology on text data along with various research areas in the field of text mining.

In Section 2.4, A review of text summarization and discussion of automatic text summarization based on various types of learning techniques used in machine learning.

In Section 2.5, Generic discussion of existing researches in the field of email summarization and implementation of various techniques have been discussed.

In Section 2.6, Detailed review on Clustering and discussion on various types of clustering algorithms which have been used for email summarization up-til now.

In Section 2.7, A summary on the literature review, the limitations and the gaps that are associated with appropriate researches based on text summarization have been outlined.

2.2 Data Mining Frameworks and ML Learning techniques

There are different frameworks that outlines the entire process used to deliver a data mining project like CRISP-DM and SEMMA.

The SEMMA process is a methodology which consists of various sequential steps like sample, explore, modify, model and assess, it was developed by the SAS Institute. [Azevedo, Santos, A. (2008)] noted that SEMMA methodology lacks business focus unlike the other methodologies. It is well observed that gathering domain and problem knowledge is not considered as any phase in SEMMA framework and its proved that while starting with a new project, without understanding the domain and the problem it is not feasible for the implementation and therefore it is assumed that these are included in sample phase of SEMMA framework. It is stated in [Azevedo, Santos, A. (2008, p. 5)] "we can integrate the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user, on the Sample stage of SEMMA, because the data cannot be sampled unless there exists a true understanding of all the presented aspects".

The CRISP-DM or Cross Industry Process for Data Mining is a framework which is well common in execution of a data mining project. The CRISP-DM [Wirth Hipp, (2000)] presents six phases of the data mining process which acts as the core steps like business understanding, data understanding, data preparation, modelling, evaluation and deployment. The Figure 1.1 highlight the iterative nature of a data mining project with its arrows pointing and connecting different steps, where the insights and results of one step may lead to manipulation of the previous or future steps for improving the quality of the project.

SEMMA	CRISP-DM
---------	Business Understanding
Sample	
Explore	Data Understanding
Modify	Data preparation
Model	Modelling
Assessment	Evaluation
---------	Deployment

Table 2.1: SEMMA and CRISP-DM frameworks (Source : Azevedo Santos, 2008)

In the knowledge discovery stage used in Machine learning, the past researches show that there are three forms of primary learning techniques like supervised learning, unsupervised learning and reinforcement learning.

In supervised learning, when the data is trained to the model or algorithm, it has respective known class labels which are considered as the equivalent correct outcomes. During the normal execution of Supervised learning technique, it generates a function that models the data using historic data records and that function is then used to predict the outcome instances of an unseen set of data. There are a lot of real world implementations using supervised learning like cancer prediction in health services [Shipp, Ross, Tamayo, et.al. (2002)],Voice recognition [Hinton, Deng, Yu, et.al., (2012)] etc. It is widely used irrespective of the type of data like numerical, text, image, mp3 files.

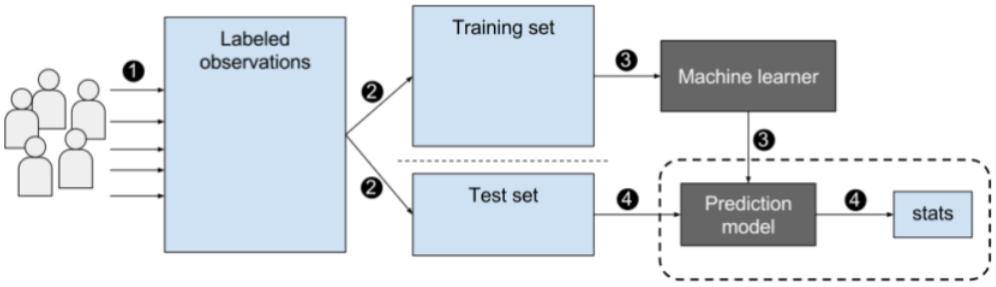


Figure 2.1: Implementation of Supervised learning

(Source: https://www.chessprogramming.org/Supervised_Learning)

The supervised learning technique is used mainly in two areas : classification problems and regression problems. This is typically done in the context of classification, when we want to map input to output labels(non-continuous/categorical outcome label), or regression, when we want to map input to a continuous output (continuous outcome label). There are various common algorithms which is used as part of the supervised leaning technique which include logistic regression, support vector machines, naive bayes, random forests and artificial neural networks. In both classification and regression, the primary goal is to determine the relationship between input variables that helps to predict the correct output class variable. Noisy or incorrect/imbalanced data labels will evidently reduce the precision of

the model and also it might lead to bias.

The next form of machine learning technique is Reinforcement learning [Barto, J. Sutton, (1997)]. This learning technique explains what to do and helps to determine the mapping situations with respective necessary actions. In addition to that, the learner is not taught what to do but instead it must learn what action gives the maximum reward with higher precise outcomes. A well common scenario of reinforcement leaning are Video games in which we complete a level and earn a badge. Defeat the bad guy in a certain number of moves and earn a bonus (Super Mario or Pac Man). If stepping into a trap them the game will get over. An example of a reinforcement learning technique is Q-learning [Watkins, (1992)], in which the goal is to reach the stage with the maximum reward and if the learner reaches at the goal/final stage, it will remain there until another learner beat that score and becomes the highest scorer.

The next form of machine learning technique is Reinforcement learning [Barto, J. Sutton, (1997)]. This learning technique explains what to do and helps to determine the mapping situations with respective necessary actions. In addition to that, the learner is not taught what to do but instead it must learn what action gives the maximum reward with higher precise outcomes. A well common scenario of reinforcement leaning are Video games in which we complete a level and earn a badge. Defeat the bad guy in a certain number of moves and earn a bonus (Super Mario or Pac Man). If stepping into a trap them the game will get over. An example of a reinforcement learning technique is Q-learning [Watkins, (1992)], in which the goal is to reach the stage with the maximum reward and if the learner reaches at the goal/final stage, it will remain there until another learner beat that score and becomes the highest scorer.

2.3 Introduction of Machine learning in text data

Data mining and Machine learning has observed rapid growth in recent years due to the advancement of technology which has led to generation and availability of various types of data [Han, J., Kamber, M. (2005)]. This is extremely true incase of text data, where the development software platforms like web and social networks has enabled a huge impact

as well as rapid creation of tremendous number of repositories of different types of data. In particular , web is a technology which enabled the creation and transfer of large amount of data in the form of textual content by various users which is in easy form to store and process. The increased amount of textual data generated and available from various applications and software has led to the advancement in algorithmic designs which can learn to determine interesting patterns and features from different types of data.

The structured data can be normally managed using database systems while textual data is which is in the form of unstructured data has to be managed using some search engines due to lack of structures in the data [Croft, W., Metzler, D., et.al (2009)]. Th search engines helps users to identify the useful information and the pattern from collection of texts using queries of different keywords. There have been several researches which was conducted in order to find the solutions to improve the effectiveness as well as the efficiency of search engines in the field of Information retrieval [Jones, K., Willett, P. (1997)]. Along with these researches many related topics like text clustering, text summarization, text categorization and recommendation systems were also studied to improve the efficiency of search engines [Grossman, D., Frieder, O. (2004)].

However, most of the researches and experiments in the field of Information retrieval was usually focused on improving information access rather than studying data to discover and identify patterns and information from those, which is the important goal of text mining. Text mining has to be focused as much important beyond the information access in order to help its users identify and analyze the information which facilitate decision making and this will also help users to analyze interesting patterns from the data, including the trends and features. Moreover, text data was analyzed in some researches depending on different levels of its representation. These text data were considered as bag of words or string of words and it was used to convert into semantic form for meaningful analysis and mining.

Unfortunately, these states of arts methods used in NLP are not still strong enough to run in unrestricted domains of textual data to generate precise semantic representations. If we consider special areas like in biomedical field and some special text mining tasks like knowledge extraction from Web, the NLP techniques particularly Information extraction are having vital role in yielding more meaningful semantic representation of text data [Pan, S.,

Yang, Q. (2010)].

There has been some recent surveys and researches in which they have identified the rapid growth of data in the context to social media which primarily falls in the area of multimedia or heterogeneous data field. [Grossman, D., Frieder, O. (2004)] Hence this has led to design of different techniques and algorithms depending on the context to different kinds of text data domains. The below mentioned are the recent research areas associated with text mining depending on different domains :

1. Information Extraction from Text Data
2. Text Summarization
3. Unsupervised Learning Methods from Text Data
4. LSI and Dimensionality Reduction for Text Mining
5. Supervised Learning Methods for Text Data
6. Transfer Learning with Text Data
7. Probabilistic Techniques for Text Mining
8. Mining Text Streams
9. Cross-Lingual Mining of Text Data
10. Text Mining in Multimedia Networks
11. Text Mining in Social Media
12. Opinion Mining from Text Data
13. Text Mining from Biomedical Data

2.4 Text Summarization

Design of Summarization systems were an interesting field of research from 1950s. Automatic text summarization is used to reduce the volume of content by generating summary from one or more paragraphs of information or from one or more text documents [Lee, J., Park, S., (2009)]. The focus of summarization can be considered as generic, which generates the important semantic information or concepts from the texts or documents, or it can be query based which can obtain the sub concepts for the user query and gives personalized abstracts depending on the match between documents, texts collection an the user input query. At Present, text summarization has gained huge research interest due to generation and the availability of vast amount of information in various forms like large text, image, audio files, video files etc. Text Summarization generates a summary of a large text that allows users to understand, exclude or include information without browsing the full text. This method of summarization can also useful in the study of classification of texts, answering questions, Information retrieval etc. This method is also implemented in several search engines especially Google search engine use the summarization technique for enhancing the search quality.

Automatic text summarization can be primarily partitioned into two methods like Unsupervised and Supervised methods , and technically Summarization methods are considered as two types into extractive and abstractive summarization. Extractive summarization is used to extract the substantial sentences using a given texts or from a set of documents and it creates the summary while in abstractive summarization, it modifies the sentence structures and might lead in the loss of meaning of sentences. An early work [Edmundson, H., (1969)] set the path for the study of applying various machine learning techniques in text summarization field of research. His work suggested that instead of depending on the single representations of topics as inputs, many other distinct topics which are important can be merged and used. Hence, a group of inputs and summaries crafted manually can be utilized to find the weight of every indicator.

In supervised learning for text summarization, the task of identifying the important sentences, partitioning the input sentences into summary and non-summary sentences are basi-

cally considered as binary classification problem [Ulrich, J. ,Murray, G. (2008)]. A corpus of manual interpretations of sentences that should be considered in the summary is utilized to train a statistical classifier algorithm and hence each sentence will be characterized into a list of likely indicators of significance. The confidence of the classifier algorithm that the sentence would be in the summary or the probability of a sentence to be a part of the summary, will be the calculated score of the sentence and this selected classifier algorithm acts the part of a sentence scoring function in which an intermediate representation of the sentence is taken as the input and the determined score of the sentences will be the output. The sentences with highest scores are selected to be a part of the summary and some of those are omitted due to the high similarity with those sentences which are already considered. The main problems associated with the supervised learning techniques are the requirement of those labelled data on which the algorithms are to be trained.

2.5 Email Summarization

Summarization must be responsive to the distinctive attributes of e-mail, a different semantic form that has the characteristics of both written text and spoken discussion. An e-mail thread or e-mailbox includes multiple conversations over time between two or more participants. Therefore, as similar to summarization of a spoken dialog, summarization must consider the interactive nature of the dialogue that is happening between participants and the response is only significant in relation to the statement it addresses. However, unlike spoken dialogue, an automatic summary does not need to concern itself with errors in speech recognition, the pronunciation impact or the availability of speech features. The responses and reactions in the case of email might not immediate and, due to the intermittent nature of the message, which specifically show the previous emails to which they are applicable.

The Study by [Nenkova, A., Bagga, A. (2003)] on summarization of email threads have explicitly implemented a summarizer using extractive summarization techniques for generating the summary for the initial two levels of the conversation thread tree, generating comparatively brief “overview summaries.” In that study they have mined a sentence from each of the two levels of the mail thread, using overlap method with previous context. An-

other later work done by [Rambow, O., Shrestha, L., et.al(2004)] on summarization of email threads have zeroed in on the dialogic nature of email. They have implemented machine learning techniques in order to create the summarizer and entrusted on giving preference to certain email features associated to the thread and features correlated to email structure like determining the count of message responders , the sentence similarities with the subject, etc along with the traditional features of summarization.

Email discussions are always a genuine way to get answers to queries and questions, and the asynchronous behavior of email discussion allows one to pursue several queries in parallel. As a result, the question-answer conversations appear to be one of the influential benefits of email discussions. These remarks led to research by [McKeown, K., Shrestha, L., et.al (2007)] on learning of question and answer sets in email and using extractive summarization techniques in machine learning for the integration of question and answer pairs of email.

The later studies by [Newman, P., Blitzer, J. (2003)] have developed an email summarizer which was designed for an entire mailbox or an archive instead of a mail thread. This approach was also designed for surfing an email from mailbox and used various multi-document summarization methods. In their study they have initially clustered all the email in related threads based on the various topics and then a full length as well as an overview summary was produced for each clusters of topics.

Email Mining can be studied as an application of Text Mining on email data for future researched and studies. Though, there are some particular characteristics of email data that set a distinct splitting line among Email mining and Text Mining:

1. Email contains further information which are included in the email headers that can be utilized for numerous email mining functionalities.

2. Texts in email are considerably briefer with less information and, hence some text mining techniques might be ineffective when using an email data.

3. Email are often briefly written and, thus, linguistic standards of writing cannot be not expected due to spelling and grammar mistakes which can appear frequently.

4. Since different topics may be discussed in a mail thread or mail chain, some of the text mining applications can be difficult e.g. mail classification.

5. Email can be private and therefore standardized methods are difficult to be implemented.

6. Email can be considered as a stream of data and ideas of target classes may alter over a period of time.

7. Email will almost certainly have noise contents; and hence those contents including the HTML tags and attachments must be eliminated so as to implement a text mining technique. In some cases, noises are intensively inserted. Especially in the case of spam emails, the noisy words and phrases are inserted, which can easily mislead to formation of ineffective machine learning models and false interpretations.

8. It is rather challenging to get a public email data for researches and studies, due to privacy related issues and GDPR establishment. This is a limitation especially for research since the studies cannot be conducted due to unavailability of public email datasets. An exemption to the beyond challenge is the Enron Corpus [Klimt, Yang, (2004)], in which this email dataset was made public after a legal investigation regarding the Enron Corporation.

2.6 Vectorised representation of words using word2vec

In recent years, Word2vec is one among the word embedding techniques that has earned the substantial attention amongst researches and studies in natural language processing. In the case of word vector, embedding vectors facilitates to discover a list of words used with respect to a given word in similar contexts. The study by [Mikolov, T., et.al (2013)] explains word embedding as an approach for expressing the meaning of a word in other words which is as demonstrated by the method of Word2vec. It makes possible to distinguish words that are used to a particular word with respect to same contexts and also helps to generate listing of words with respect to a word in the specific context. Although these Embedding has

achieved considerable interest between the NLP researchers in recent years, the potential for using these techniques in the process of information retrieval (IR) has so far been scarcely explored.

The study from [Hinton, G.,(1986)] depicts the long story of representing words terms as continuous vectors. One of the standard models has been recommended in [Bengio, Y., (2003)] to estimate the neural network language model (NNLM) in which the statistical language model and the vector representation of words were learned jointly by utilizing a feedforward neural network along with a linear projection layer and a non-linear hidden layer. The researchers from Google came up with an open source language modeling tools named Word2Vec [Mikolov, T., et.al (2013)] , which could learn from a large number of unstructured data and retrieve the semantic and syntactic information. In the field of natural language processing, this approach has gained broad interest. This method uses the deep learning concepts in order to train each word and map them into a k dimensional word vector. It helps to determine the semantic similarity from the text by measuring the cosine similarity between the word vectors in the vector space.

In order to train the word vectors swiftly and effectively, Word2Vec utilizes the training models like continuous bag-of-word model and skip-gram model [Mikolov, T., (2013)], in which the CBOW predict the words related to the current context and the skip gram helps to predict the context through the current word. Basically, this technique consider text as input set and it creates the respective word vector efficiently in the learning and training process. This method executes both the models by taking the tokenized input text normally but then also it intakes input text which are not processed and develops a feature vector. In the natural language processing, the word vector is used to determine the semantic features between the words and also it measures the cosine similarity among the words which can be utilized effectively for clustering, speech analysis etc.

Word embedding are normally learned on graphs from knowledge, while it is studied that the most common models learn these representations of vectors only by using large corpus of data. In fact, all these models are based on one theory on which the words that occur in the same contexts will have parallel likely meanings. Most of the corpus word embedding models are original Word2Vec variations. Various document embedding models are specific

word embeddings that enhance word embedding to provide significant document embedding only by averaging or doing a similar composition. Word embedding can be constructed into more abstract constructs like sentences, phrases, paragraphs etc. Most unsupervised algorithms still use a very similar model to word embedding framework encouraged by the Word2Vec algorithm.

2.7 Introduction of BERT in the field of NLP

Pre-training of the language model has been demonstrated to be enhanced for several natural language processing responsibilities [Dai, A. , Le, Q. (2018); Radford, A., et al., (2018)]. Especially important tasks like Natural language inference [Bowman, S. et al., (2015); Williams, A., et al., (2018)], paraphrasing [Dolan, W., Brockett, C., (2005)] which intend to determine the relations among the sentences by studying them comprehensively. Another intention to these tasks is also for tokenizing tasks like question answering and named entity recognition in which the models are used to generate fine output from the token level [Sang, T., Meulder, D., (2003); Rajpurkar, P. et al., (2016)].

In the field of research for downstream tasks, there are two current approaches for utilizing pre-trained language representations like feature-based and fine tuning. Among those the feature based strategies like ELMo uses various architecture which are specific to the tasks and it incorporates the pre-trained representations as additional features [Peters, M., et al (2018a)]. While in the case of fine-tuning strategies like Generative Pre-trained Transformer : OpenAI GPT [Radford, A., et al., (2018)], it proposes nominal task-specific parameters, and are specifically prepared for the downstream tasks through fine-tuning all parameters which are pretrained. During pre-training, the both these methods share the same objective purpose, using unidirectional language models for identifying and learning the language representations. It will well studied that the current techniques limit the power of pre-trained representations, particularly with regard to fine-tuning approaches.

It will well be studied that the current methods limit the quality of pre-trained representations, particularly with regard to fine-tuning strategies. The main drawback is the unidirectional nature of standard language models, which limits the choice of architectures that

can be applied through pre-training procedures. For example, [Vaswani, A., et al., (2017)] the authors make use of a left-to-right architecture in OpenAI GPT, by which each token can only interact with earlier tokens in the Transformer's self-attention layers. These limitations are sub-optimum when used in the case of sentence-level tasks and could be dangerous when implementing fine tuning-based strategies to token-level activities like answering questions where context in both directions are important to implement.

The latest approach for enhancing the fine tuning strategy was designed by Google researchers in 2018 by introducing Bidirectional Encoder Representations from Transformers. BERT is a technique for pre-training representations of language which are widely used to build models that can be reproduced and utilized free of charge by NLP researchers. This method works with "Masked Language Model" (MLM) which is a pre-training objective inspired by the Cloze task [Taylor, W., (1953).] through which BERT eases the previously stated unidirectionality limitation. The MLM model masks some of the input tokens at random, and the aim is to predict the masked word's original vocabulary id based on the context meaning alone. Contrasting to the pre-training of left-to-right model, the masked language model aims to facilitate the representation to combine the context of left and right and thus enabling to pre-train a bidirectional transformer.

There has been a long term research study regarding general language representations pre-training, and the very commonly used approaches are briefed in the below:

2.7.1 Feature-based Unsupervised Approaches

The study of Word representations that are widely applicable have been an active research area over the past, including non-neural [Brown, P., et al., (1992); Ando, R., Zhang, T., (2005)] and the neural methods [Mikolov, T., et al., (2013); Pennington, J., et al., (2014)]. One of the fundamental parts in the modern NLP is Pre-trained word embedding, and the studies offer a significant enhancement in the embeddings which are learned from the basics [Turian, J., et al., (2010)].Various objective like left-to-right language modeling [Mnih, A., Hinton, G., (2009)] and the intentions to distinguish correct words from incorrect ones in the context to left and right [Mikolov, T., et al., (2013)] has been used to pre- train the word embedding vectors. Such approach of studies is extended to grosser granularities, such as

the embedding of paragraphs [Le, Q., Mikolov, T., (2014)] and sentence embeddings [Kiros, R., et al., (2015)]. For preparing sentence representations, prior works of rank candidate next sentences [Jernite, Y., et al., (2017); Logeswaran, L., Lee, H., (2018)] and creation of next sentence left-to-right by using a representation of the words from the preceding sentence [Kiros, R., et al., (2015)].

Traditional researches of word embedding simplify ELMo and its predecessors along a diverse aspect [Peters, M., et al., (2017, 2018a)]. These approaches used to extract those sensitive features based on the context using left-to-right and a right-to-left language model and the concatenation of left-to-right words with right-to-left word representations was used for generating contextual representation for every token. It is well proved that the integration of contextual word representation embeddings with current task-specific designs, ELMo enhances as well as expands the state of the art for numerous key NLP standards and architectures [Peters, M., et al., (2017, 2018a)] like sentiment analysis [Socher, R., et al., (2013)], named entity recognition (Sang, T., De Meulder, F., 2003) and question answering [Rajpurkar, P. et al., (2016)]. Another research by [Melamud et al. (2016)] introduced LSTMs for predicting a single word from both left and right context as part of learning the contextual representation in sentences. This model using LSTMs are also feature based and are not bidirectional.

2.7.2 Fine-tuning Unsupervised Approaches

The initial works with feature based strategies was from unlabeled texts using pre-trained word embedding parameters. Recent researches focusing on sentence encoders which generate token representations based on the context are pre-trained using the unlabeled texts and are fine tuned for various downstream tasks which are supervised [Radford, A., et al., (2018)]. The benefit of these strategies is that it is necessary to learn from scratch only few parameters. In such model objectives like autoencoder and Left-to-right language modeling have been utilized for the pre training purpose [Dai, A. , Le, Q. (2018); Radford, A., et al., (2018)]. There were also much works showing much effective transfer with larger data sets for supervised tasks like machine translations and natural language inferences. BERT becomes the initial model based on fine-tuning based representation to attain state-of - the-

art performance on a broad range of tasks like sentence-level tasks as well as token-level, exceeding several task-specific designs and architectures.

The below figure explains the procedures of pre-training and fine- tuning which occurs when using BERT technique. In addition to output layers, in both pre-training and fine tuning strategies, the similar architectures are applied. Models for various downstream tasks are structured using the similar pre-trained model parameters while in the case of fine- tuning approach, each of the parameters are tuned fine. In the diagram, [CLS] is a special sign applied to each instance input and [SEP] is used as a separator token.

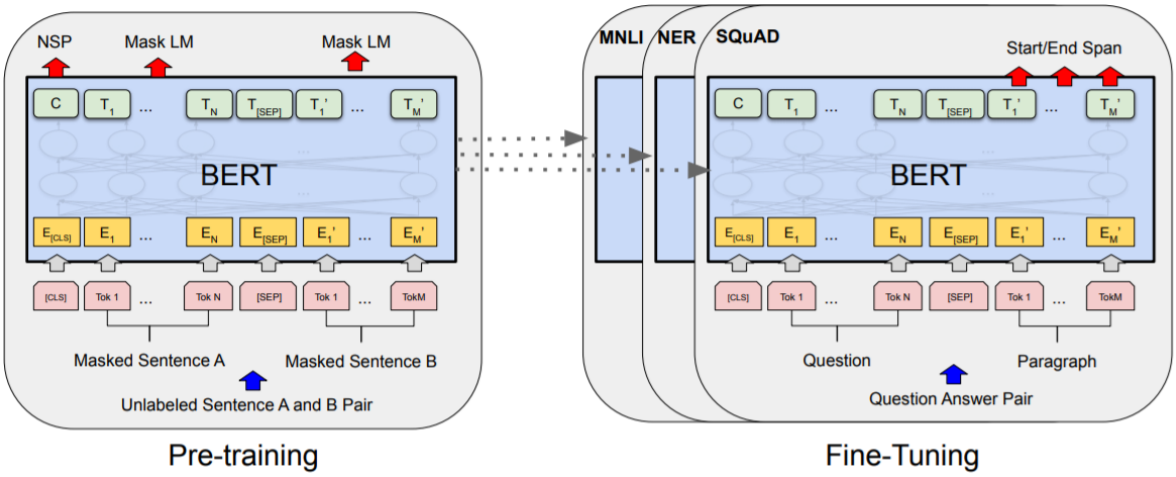


Figure 2.2: BERT:Pre-training and fine-tuning procedures (Source:Devlin, J., et al 2019)

2.7.3 BERT Framework

In the case of Bidirectional Encoder Representations from Transformers (BERT), there are majorly two steps in the framework like pre-training and fine-tuning. Initially the model is trained with the help of an unlabeled data over various pre- training tasks. The BERT model is then modified with the parameters of pre-trained step for the purpose of fine-tuning, and then every parameter is fine-tuned again with the help of labeled data for downstream purposes. Then each downstream role has different fine-tuned models, despite initializing them with the similar pre-trained parameters.

BERT's unified architecture across various tasks is a unique feature while the pre-trained architecture and the final downstream architecture are minimally different. BERT is a bidi-

rectional Transformer encoder model which is multi-layered and centered on the original study as in [Vaswani, A., et al. (2017)] and it was released globally using tensor2tensor library.[1] The basic working of transformers in NLP along with the background explanation of the model design are referred in [Vaswani, A., et al. (2017)] as well as exceptional handbooks such as "The Annotated Transformer".[2]

In order to ensure that BERT performs a number of downstream functions, the input representation should clearly represent a single as well as a pair of sentences in a sequence of one token string. A "sentence" may be a random string of contiguous text rather than a linguistic sentence and "sequence" refers to the BERT input token set, which can be a single or two sentences combined. A special classification token ([CLS]) is always the first token of each sequence. The final secret state corresponding to this token is used for classification tasks as the cumulative sequence representation. Pairs of sentences are packed into one sequence and they are distinguished in two ways. Initially they are separated with [SEP] token a learned embedding is added to all the tokens which indicate whether the sentences belong to each other.

Like [Peters et al. (2018a), Radford et al. (2018)], in order to pre-train BERT, we do not use conventional language models like left to right or right to left. Alternatively, we are pre-training BERT using unsupervised tasks mentioned below.

Task 1: Masked LM

Instinctively, this is the major reason to consider that BERT, a deep bidirectional model is precisely more effective than other models like left-to-right or the shallow concatenation of another model. To train a deep bidirectional model representation, we randomly mask a certain percentage of the input tokens and then tries to predict other tokens which are masked. This procedure is called as Masked Language Model (MLM), as in a typical LM, the final hidden vectors related to the tokens which are masked are supplied into a softmax output across the vocabulary. While this helps us to obtain a pre-trained bidirectional model, a drawback is that we establish a discrepancy between pre-training and fine tuning, as the [MASK] token does not appear during fine tuning.

[1] https://github.com/tensorflow/tensor2tensor

[2] http://nlp.seas.harvard.edu/2018/04/03/attention.html

Task 2: Next Sentence Prediction (NSP)

Most of the key tasks in downstream processes like natural language inference and answering questions are all on the basis of the relation between two sentences and understanding those relationship, but this is not captures directly using a language model. To train a model which understands the relationships of sentences, we are pre-training for a binarized task of next sentence prediction that can be created slightly from any corpus. The NSP task is directly linked to representation discovering ideas used in [Jernite, Y., et al. (2017)]. In previous work, however, only sentence embedding is passed to downstream tasks, where BERT shifts all parameters to modify model parameters for the final task. The pre-training procedure follows mostly the current literature on the pre-training language model.

2.8 Clustering – An unsupervised learning technique

Clustering is an extensively implemented machine learning technique in research field of the text summarization and other domains. The challenges associated with this technique are in numerous NLP applications like classification of text data, customer segmentation, collaborative filtering, document organization, text visualization, text indexing etc. This Unsupervised learning technique is applied for identifying the groups of similar data objects and the similarity of these data objects are measured using similarity function. It is widely used in text domains where the data objects to be clustered and grouped can be from various text representations such as documents, paragraphs, sentences, words or terms. Hence it makes a huge impact from organizing documents to enhance information retrieval.

Traditional research methods of implementing clustering have mostly centered around the quantitative data [Ng, R., Han, J. (1994)] , in which the attributes or the data objects are numeric. There have been various text summarization researches which was implemented by using an efficient clustering process, where the frequency of words which are standardized in terms of its relative frequency of presence in the text/document and throughout the entire collection was used. Normally, in text processing a common representation of vector space based TF-IDF are used [Salton, G. (1983)]. In this representation of TF-IDF, the term frequency of all the words in the text/document is standardized by implementing IDF

or inverse document frequency. Inverse document frequency standardization decreases the weight of the terms that occur more repeatedly in the collection which help in avoiding the recurrence of words. It decreases the importance of common terms in the collection, meaning that the matching of texts/documents is more affected by more discriminative words of relatively less frequency in the collection. Along with IDF, a sublinear transformation function is regularly used to the term with highest frequencies in order to prevent the adverse prevailing effect of any single term that are common in a document.

The algorithms used in Text clustering are classified into a wide range of various forms, such as agglomerative clustering algorithms, partitioning algorithms, and regular parametric modeling algorithms. These various representations require the design of different clustering algorithm classes. In terms of efficiency and quality, different clustering algorithms have different tradeoffs.

2.8.1 Feature Selection and Transformation Methods

Irrespective of learning type, whether its classification or clustering, the quality and the reliability of any machine learning methods are extremely dependent on the noisiness or the variations of the features that are used for the clustering process. If we consider simple example, commonly used words in texts such as “the”, “this” etc. might not be helpful in enhancing the quality of clustering. Thus, during feature selection, it is always crucial to determine and choose the features from the texts efficiently, so that the words which are not valuable or might create noise in the corpus are eliminated before the clustering algorithms are executed. In addition to feature selection its always preferred to implement other feature transformation methods, so as to enhance the quality of the document representation and render it more flexible to clustering algorithms. Various feature transformation methods available are Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Analysis (PLSA), and Non-negative Matrix Factorization (NMF). In all these techniques which are used for feature transformation, the correlations amongst the words in the lexicon/texts are leveraged and determined so as to generate features among them, which resemble to the concepts and the principal components of the text data.

2.8.2 Agglomerative and Hierarchical Clustering Algorithms

The study of Hierarchical clustering algorithms was widely performed during the past researches in clustering literature. These algorithms were mostly designed and implemented with in the context of text data field of studies. [Jardine, N., Rijsbergen, J. (1971), Jain, A., Dubes, R. (1998)] explains the various implementation uses of these algorithms for different kind of data including numerical quantitative data, categorical data and unstructured text data. The study by [Murtagh, F. (1983)]provides an outline about conventional hierarchical and agglomerative clustering algorithms in the perspective of unstructured text data. The agglomerative hierarchical clustering approach is particularly useful to help a number of searching techniques, because it automatically generates a tree-like hierarchy that can be used for the efficient search process and for enhancing the searching capability of algorithms. In specific, the efficacy of these algorithms in enhancing the search effectiveness over a subsequent scan has been discussed in [Jardine, N., Rijsbergen, J. (1971)].

The common concept in implementation of agglomerative clustering is to sequentially merge and combine the texts or documents into various clusters on the basis of their similarity to each.Computing text similarity is a key issue in the field of information retrieval from text data. Most of the researches have shown that all hierarchical clustering algorithms continuously join groups on the basis of the most appropriate pairwise similarity amongst those sets of texts or documents. The major variations between these types of approaches are the manner in which this pairwise correlation/similarity is assessed between various groups or sets of texts. If we consider an example, the similarity among a couple of groups may be measured as the best-case, the average case, or the worst-case similarity amongst the texts/documents taken from these sets of groups. Such similarity functions can be implemented in tandem with a broad range of conventional clustering algorithms. Theoretically, the method of agglomerating texts to sequentially higher cluster levels produces a cluster hierarchy for which the leaf nodes resemble to individual texts/documents, and the inner nodes relate to those cluster groups which are merged. If two groups are combined, a new node corresponding to this larger incorporated group is formed in this tree. The two children of this newly formed node refer to the two groups or classes of documents which are combined into it.

The various techniques for combining groups or sets of texts/documents with several agglomerative approaches are as follows.

- Single Linkage Clustering:

  In the case of single linkage clustering, it is discovered that the similarity among two sets of documents is the highest similarity amongst any pair of documents of that group. In this technique, the two groups are merged in such a way that the nearby pair of documents are supposed to have the greatest similarity when compared to any of the other combinations of pairs of such groups. The major advantage of using single linkage clustering is that they are highly effective when applied. This is because we can initially compute all pairs based in their similarities and then sort it descending order of similarities. These pairs are considered in a pre-defined order and they are merged successively if the pairs are from different groups. This is ultimately parallel to a bridging tree algorithm based upon the pairwise-distances by managing the fringes of the graph in a specific order.

  The major problem of this method is that it can cause a chaining effect in which a sequence of identical documents causes to different documents being clustered into the same clusters. This clustering facilitates the grouping of texts through such chains of transitivity. This can lead to formation of poor clusters, mainly when higher orders of agglomerative clusters are implemented.

- Group-Average Linkage Clustering:

  In the case of group-average linkage clustering technique, the similarity among two clusters is computed as the average or mean similarity among the sets of documents or texts in the two clusters. It is studies that the average link clustering method is denser than single-link clustering, because the average of the similarities is computed between a significant number of sets so as to decide the group-wise similarity. But this technique is much more robust when considering the quality of the clusters, because it does not show behavior of chaining effect as of single linkage clustering.

  Various researched have displayed that the performance of the average linkage clustering algorithm can be increased by approximating the average linkage similarity

between two clusters. But this approach will work effectively only in text data domain rather than all other data domains. In this scenario, the execution time of group-average linkage clustering technique can be lowered to O(n2), where n depicts the total count of nodes. This technique can be applied very effectively in the case of document or text data, since a cluster centroid will be the concatenation of the documents or texts in that cluster.

- Complete Linkage Clustering:

  In the case of complete linkage clustering algorithm, the similarity between the two clusters is measured as the worst-case relation between any document pair of the two clusters. This method of clustering also helps to avoid chaining effect because it prevents the positioning of any pair of extremely different points in the same cluster.

  Though, similar to group average clustering, it is much complex than the single-linkage clustering method. The complete linkage clustering technique requires O(n2) space and O(n3) time as compared with other linkage clustering methods. However, the space capacity of complete linkage clustering can be significantly reduced when using text data, this is because the most of pairwise similarities will be zero in the case of complete linkage clustering algorithm.

2.8.3 Distance-based Partitioning Algorithms

Distance based Partitioning algorithms are commonly used in database literature to efficiently construct object clusters. The two most commonly used distance-based partitioning algorithms are the following:

- k-medoid clustering algorithms:

  In the case of k-medoid clustering, normally a set of points from the data are used as medoids or anchors on which the clusters are developed. The major aim of this clustering algorithm is to define an ideal group of representative texts from the initial set of data corpus over which the clusters are developed and each document/text are allocated to its nearest representative neighbor from the set of collection of data.

This develops a running group of clusters from the data corpus which are sequentially enhanced by a randomized method. This clustering algorithm operates with an iterative execution in which the group of k representatives are sequentially advanced with the help of randomized inter-changes from the data. In Particular, the group-average similarity is executed with each text/ document in the data corpus to its nearest representative as the actual function which needs to be enhanced during this process. During each repetition, the arbitrarily picked representative are replaced and changed in the existing group of medoids with a randomly selected object from the data collection, if it enhances the clustering objective functionality. This technique is iterated till the convergence criteria is attained.

There are two major drawbacks of the using the k-medoids based clustering algorithms which are specific to the case of text data domains. The common drawback of k-medoids clustering algorithms is that this need a larger count of repetitions in order to attain convergence and hence the execution are very slow. This is basically due to each repetition that involves the calculation of an objective task whose time constraint is proportionate to the magnitude of the main data corpus. Another disadvantage of using the k-medoids based clustering algorithms is that this algorithm is not well efficient when its executed with some data such as text. This is due to the enormous amount of document pairs that lack the words which are in common and due to the lack of similarities between the pairs and larger impact of the noise. This disadvantage is certain only in the case of the textual data domain and information retrieval domain, due to its sparse nature of the fundamental text data and information retrieval.

- k-means clustering algorithms:

  The k-means clustering algorithm is another kind of Distance based Partitioning algorithms which also uses a ‘k’ value around which the clusters are developed. Though, this k value for representatives are not certainly attained from the original data and are enhanced in a different way than k-medoids technique. The easiest way of the k-means technique is to begin the iteration with a set of k values from the initial corpus

and assign data objects to various clusters on the basis of nearest similarity. During the next iteration, the centroid of the assigned clusters to each object are used to replace the centroid for data objects in the last iteration. In other words, the new set of data objects are defined, so that it is the next centroid for this cluster based on the object similarity and this approach of formation of new clusters are continued until convergence criteria is met.

In some ways k-means algorithm is more advantageous than the k-medoids algorithm, and one of the main advantages are that k-means algorithm needs very a smaller number of iterations in order to meet the convergence criteria. The study from [Schutze, H., Silverstein, C., (1997)] recommend that for several large datasets, it is enough to apply 5 or fewer iterations for an efficient clustering.

The major disadvantage of the k-means algorithm is that it is extremely sensitive to the set of data clusters which was formed during the clustering in the initial iteration. In addition, due to the occurrence of large number of words around the given cluster of texts will slow the similarity computation for the next iterations.

The existing researches in the field of information retrieval systems majorly comprises of various phases like Retrieval phase followed by Clustering and Summarization phase. QCS [Dunlavy, D., O'Leary, D., et.al (2007)], which is used for querying, clustering and summarizing the documents, is basically an information retrieval system that incorporates all the above three phases. In the initial querying phase of this system, it retrieves a group of those documents which are relevant when given through an input query using Latent Semantic Indexing (LSI) method. While in clustering part, those documents which are retrieved in initial phase are clustered into various groups by spherical k- means algorithm and a summary is being generated from all those clusters which was formed in Clustering phase with the help of pivoted QR decomposition method and Hidden Markov Model (HMM) as part of summarization phase. However, the value of 'k' used for the k-means algorithm in clustering phase is given as user input. The clustering part in those systems are used to retrieve documents/Texts as per the topic groups and it enhances the information retrieval system by delivering the focused information in organized form. The summarization part delivers an

abstract of large documents and helps the users to quickly search the important information which are relevant without going through the entire text.

In the research field of Information retrieval, Hierarchical Agglomerative Clustering (HAC) methods are widely used in order to enhance the efficiency the systems. [Tombros, A., Villa, R., et.al (2002)] suggested a suitable method for enhancing the effectiveness of Information retrieval systems. That proposed system utilizes this HAC method for clustering and it is specifically used in query based clustering. However, the clustering methods which are used in these systems are influenced by various disadvantages or problems like :

- Chaining effect
- Convergence criteria for HAC method
- Deciding the number of clusters (k-value)

2.9 Summary, Limitations and Gaps in Literature Survey

Text summarization is a fairly well-studied problem in literature right from the late 1950s. One of the first attempts to solve this problem came from [Luhn,H.(1958)] which used high-frequency words present in the document to score a sentence for relevance. Over the years, several techniques have been applied for solving this problem including some recent attempts using neural networks [(Lu, Z., Li, H., 2013) (Kupiec, J., Pedersen, J. and Chen, F., 1995) (Turney, P.D., 2000)]. [(Luhn, H. ,1958) (Yang, Y., Liu, X. ,1999) (Hovy, E., Lin, C. ,1997) (Wang, H., Lu, Z., Li, H., Chen, E., 2013)(Wan, X., 2007)] have used various forms of attention based encoder and decoder models to generate keyword/headline style summaries. In contrast, this method is used for generating multi-sentence summaries, where sentences in the summary are expected to deal with distinct semantic concepts. [(Voorhees, E. M., 2002) (Zhai, C., Lafferty, J., 2001) (Berger, A.L and Mittal, V.O, 2000)] use an auto-encoder to learn a low dimensional embedding of a paragraph and could be potentially be used for summarization because as the length of the content in the paragraph increases, the system is likely to generate a condensed version of the original paragraph which is partially retrieving the exact meaning or summary of it. Early research on extractive summarization

is based on simple heuristic features of the sentences such as their position in the text, the overall frequency of the words they contain, or some key phrases indicating the importance of the sentences [(Lewis, D. D., Knowles, K. A. ,1997) (Scheffer, T. ,2004) (Segal, R. B., Kephart, J. O. ,1999) (Scheffer, T., 2004)].

The idea of clustering sentences in a high dimensional space has also been used for text summarization in the past. However, those systems used TFIDF representations of sentences (which are only applicable in a multidocument summarization system) instead of sentence embeddings. A recent research by [Padmakumar ,A., Saran, A., (2016)] using tipster 2 dataset containing 183 documents, in which various different types of clustering algorithms like k- means, mean shift was used indivdually in order to obtain the text summarization and a maximum precision rate of 0.41 for k- means clustering was obtained when evaluated using ROUGE method . Another research by [Naveen, G., Nedungadi, P., (2014)] explained that Hierarchical Agglomerative Clustering (HAC) methods have been widely applied in the field of Information Retrieval and summarization as the techniques which can improve the efficiency as well as the performance of the Information Retrieval systems. However, clustering techniques used in these kind of information retrieval and summarization systems are affected by numerous disadvantages like:

a. Chaining effect which is a common problem in Single Linkage clustering and PHA method and it can be defined as the gradual growth of a single cluster as one data object with the elements added to that cluster at each iterative step of the algorithm.

b. Deciding optimum number of clusters k is the major problem with Partitioning-based clustering approaches such as k-means method.

The proposed hybrid clustering method for clustering solves these disadvantages. In the research work [Naveen, G., Nedungadi, P., (2014)] it was concluded that using PHA-ClusteringGain-KMeans hybrid clustering approach, by combining PHA method and k-means method is more efficient and accurate than conventional Hierarchical Agglomerative Clustering (HAC) algorithm. However, an assessment on the level to which a hybrid model can out-perform comparatively to a conventional k-means clustering algorithm in automatic text summarization has not been discussed. In addition to that, as we know feature embedding techniques which are vital in generating sentence/ feature vectors and these vectors are

getting involved in clustering and the modelling purpose but there has not been any discussion on any impact of feature embedding techniques in the process of text summarization. Hence, through this research a comparative study is being performed by implementing the text summarization task using different approaches such as hybrid/conventional Unsupervised learning algorithms along with various combinations of feature embedding techniques like Word2Vec/BERT and also to determine the finest approach that outperforms in terms of precision.

To address the discussed limitations and research gaps presented in this section, the research question is given as:

"To what extent can a PHA-Clustering Gain k-Means hybrid approach out-perform a conventional k -means clustering model for email-based text summarization ?"

The next sections will discuss in detail, the research design, implementation and evaluation of experiments to address the research question.

Chapter 3

Experiment design and methodology

3.1 Introduction

This part introduces the detailed information based on the design and methodology which are applied in the research to solve the research question which is discussed in Chapter 1. The Cross Industry Standard Process for Data Mining (CRISP - DM) methodology is practiced in the research lifecycle of this study. The experiments in this research is carried out in Jupyter Notebooks using Python programming.

Section 3.2 outlines the Business Understanding of this research which explains an analysis of the business objectives in terms of problem definition and needs of this research. Present situation is studied and using these understandings, the goals of carrying out the methods are briefed.

Section 3.3 describes the Data understanding by which the initial facts and figures collection are done from all available sources and the properties of the data used in this research is analyzed. Then the quality of data is validated which are related to the completeness and accuracy of data.

Section 3.4 explains the Data preparation and various procedures associated with pre–processing of data. The data is completely explored and data sampling, data cleaning and feature engi-

neering of data is performed in this part. Explored information from this phase are utilized to identify the patterns in light of business understandings.

Section 3.5 explains the Modelling phase, in which the selection of modelling techniques followed by the generation of test scenario for validating the model's quality are implemented and all the models are then assessed to make sure that they fall in line with the business objectives.

Section 3.6 discuss the evaluation phase, proper evaluation technique is chosen and using this technique the results of models which are achieved from the previous section are evaluated in the backdrop of business intentions.

The workflow of the experiments in terms of CRISP- DM research life cycle is as shown below:

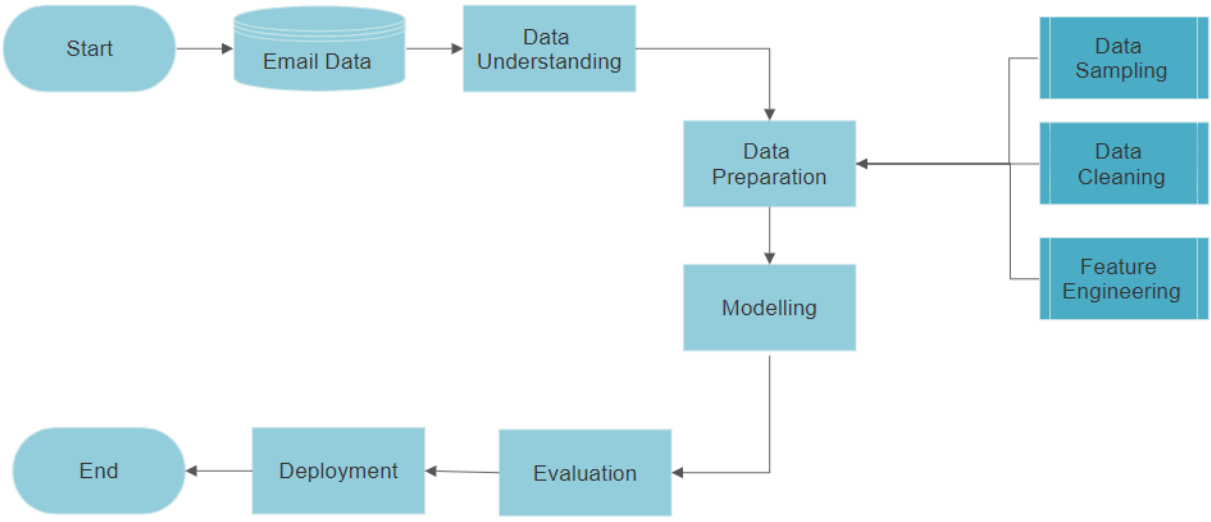


Figure 3.1: Flow chart in terms of CRISP-DM methodology

3.2 Business understanding

The impact of electronic mail is now more evident than ever in our daily lives. Millions of plain text or enriched messages are sent and received around the globe every minute. Some of them are read with extra care and at the same time, with obvious disinterest, many of them are deleted. As the internet grows, electronic mail has become not only a vital tool for our work, but also an important means of communication between people.

Email also significantly facilitated personal communication as it offered instant messaging with minimal cost. People from around the world can now exchange views and information so easily that email has become the second most popular channel of voice after voice communications. Features that made email so popular are the speed of communication, the minimum cost, and the remarkably easy use of it. An advantage over voice communication (e.g. telephone) is that it is asynchronous, meaning that both sides of communication do not need to be simultaneously online or in front of a computer. Unfortunately, the curse of Information Overload could not escape email. Loads and loads of incoming messages have turned the handling of electronic mail into a tedious task (some extremely important, others simply junk). So, with the application of Text summarization on electronic mail, we can extensively save the time as well as retrieve the precised meaning without losing any information from the main content.

The text mining technique like text summarization for large texts can be better addressed using hybrid approach of clustering compared to conventional approaches. The main motive of the research is to determine performance and compare precision of a hybrid approach along with a conventional approach of clustering in text summarization. The hypothesis of this research is as follows:

H0 :The application of PHA-ClusteringGain k-Means hybrid approach in text summarization will result in no precision increase over a conventional k-means clustering model.

H1 :The application of PHA-ClusteringGain k-Means hybrid approach will lead to an increase of precision rate in text summarization over a conventional k-means clustering model.

The following diagram explains the experimental design which is applied in this research :

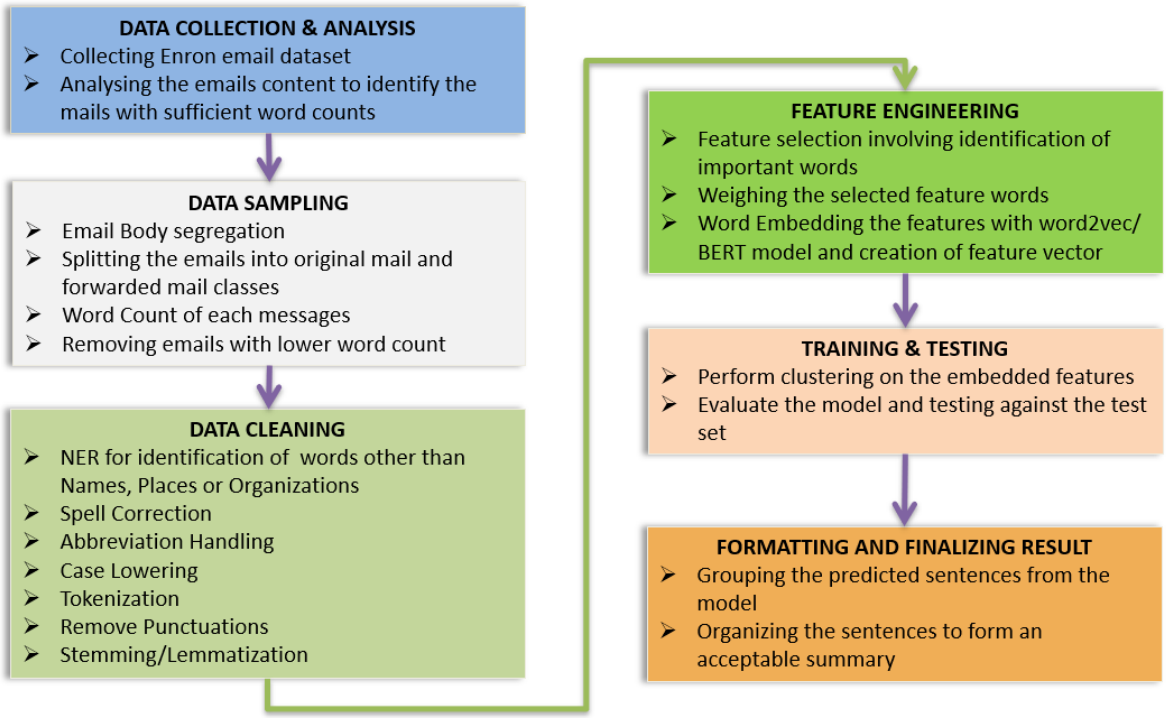


Figure 3.2: Experimental design and process flow diagram

3.3 Data Understanding

One of the main limitation of this study was challenging to get a public email data for researches and studies, due to privacy related issues and GDPR establishment. This is a drawback and a limitation especially for research since the studies cannot be conducted due to unavailability of public email datasets.

An exemption to the beyond challenge is the Enron Corpus data [Klimt,Yang, (2004)], in which this email data was made public after a legal investigation regarding the Enron Corporation.

Link : https://www.kaggle.com/wcukierski/enronemail-dataset

The Enron email dataset provides real-world text data that is arguably of the same kind as data from Echelon intercepts – a set of messages about a wide range of topics, from a large group of people. The Enron email dataset contains electronic mails in text format generated by employees of the Enron Corporation. It was obtained by the Federal Energy Regulatory Commission during its investigation of Enron's collapse.

This dataset contains 517,401 observations with 2 features/variables both in text type data.

The features/variables of the data are discussed in the following table:

Attribute	**Description**	**Type**
file	File content which contains filename	Text
message	Message content with multiple information including the actual message body	Text

Table 3.1: Description of features/variables of the dataset

More detailed understanding of the data is performed using various analytical methods and python libraries like pandas and matplotlib :

1. Finding total number of records from the dataset.
2. Determining the number of duplicate emails.
3. Finding the emails with no data in the message body and removing those.
4. Analysing the email content to identify the mails with sufficient word counts.A benchmark word count was determined using statistical analysis and for text summarization purpose, only those emails with a minimum count of (329) words are considered for processing.
5. Checking those messages which are not having X-FileName since that is our splitting benchmark.
6. Determining those emails which are forwarded and removing those from train data so as to avoid the discrepancy and bias in the data.

7. Splitting the entire data into two sets ”messages and forwarded messages” based on the presence of forwarded information.

3.4 Data preparation

In this phase of data preparation, necessary steps are performed before processing the data for modeling phase. It is important to consider this phase to improve the quality of models and to avoid the bias in the model. This phase of data preparation includes Data sampling, Data cleaning and followed with feature engineering.

3.4.1 Data Sampling

This is an important step that has to be carried with this data especially Enron email data. The salutation and signature lines at the beginning and the end of an email contributes no value for the task of summary generation. So, it is necessary to remove these lines from the email, which we know, should not contribute to the summary.

While doing data understanding analysis, the features was in text type and the main feature named ‘message’ was a combination of a lot of information and it is vital to segregate the email message from all other information. Another important task was designed to split the emails into original email and forwarded mail class in order to avoid the discrepancy in the data and to remove the bias in the models. Determining the word count is another part of this phase where analyzing the email content to identify the mails with sufficient word counts. As a benchmark mean word count was determined and for text summarization purpose only those emails with a minimum count of mean word count (329) words are considered for processing. Hence, emails with lower word count are removed from the data.

3.4.2 Data Cleaning

Data cleaning is the method of identifying and correcting inaccurate information from a record set, data set and it refers to the detection of missing, wrong, corrupted or irrelevant parts of the data and the substitution, alteration or removal of raw data. Data cleaning is

not just a matter of removing information to make room for new data, but it is a rather of discovering a way to maximize the accuracy of the data set without necessarily removing any information content from the data. Data cleaning helps to ensure that information still fits the right fields while making it easier for data models to connect with data sets and find information more effectively.

Various processes involved in this phase are as follows:

1. Spell Correction

   It is important to do the spell correction when using a text data. There are various libraries which are widely available in python for doing this process. For this implementation different python libraries like ‘spacy’ and ‘re’ are used for this task. We can make a very common spellchecker with the help of a dictionary lookup as methods in Python. Spell check correction on individual words from the sentence or texts can also be done using other packages like ’enchant’ in Python. But There are some improved string algorithms in spacy and re libraries that have been developed to fit the fuzzy sequence.

2. NER for identification of words other than Names, Places or Organizations.

   Named Entity Recognition is an NLP technique in which an algorithm takes the text or sentences as a string of input text and it recognizes the related nouns like people, places, and organizations that are listed in that string of text or sentences. This technique is well commonly used for the analysis of large articles in which it scans the entire articles or paragraphs automatically and identifies all nouns which are discussed in it. This process also helps in categorizing the articles in a hierarchical approach and helps in discovery on contents from it.

3. Abbreviation Handling

   In a text of conversation or chats, people tend to use various abbreviations while writing or typing and it is well common. While data processing it is vital to correct these abbreviations, which are used and reform the sentences in the actual way it is supposed to be. This was implemented using an external file named ’chatshortform.xlsx,

in which all the abbreviations and its expansions are being mentioned in this file. The basic task is to identify the abbreviations from each sentence and replace those with the expansion which are available from the abbreviation file.

4. Case Lowering and Remove Punctuations

   From a paragraph of texts there are some chances for the mis use of lower case and upper case, in this all the sentences are being reformed into lower cases and all the punctuations are removed.

5. Tokenization

   This is a technique which are used in order to split or tokenize text into list of tokens. Tokens are said to be word in a sentence and a sentence in a paragraph. These tokenizers basically separated the words using spaces and punctuations. Once the languages identification is performed for every email, we can use this information to split each email into its constituent sentences using specific rules for sentence delimiters for each language. NLTK Tokenizer package in Python helps to divide strings into lists of substrings. It can be used to find the words and punctuation in a string and this is basically the initial step to be performed in NLP as part of prep-rocessing. This process consists of three phases like dividing the sentences into words, realizing the significance of each word with reverence to the respective sentence and ultimately generate structural explanation on the input sentence.

6. Stemming/Lemmatization

   Stemming and Lemmatization are Text Normalization process or referred to as Word Normalization methods in the area of NLP which are used primarily to organize texts, words and sentences in documents for advanced processing. xStemming algorithms are empirical methods work by cutting the end or beginning of a word in the hope of the goal correctly, considering a list of mutual prefixes and suffixes that can be contained in an inflected term and words. Porter's stemming algorithm is a well common algorithm which is used to perform stemming of words from the sentence. Stemming can reduce the memory space required to store the words and makes computation eas-

ier. While Lemmatization, on the other hand, considers the morphological study of words by which it is important to have comprehensive dictionaries that the algorithm can consider through to connect the form back to its lemma factor.

3.4.3 Feature engineering

Feature engineering is the most crucial part while developing NLP studies. Features are acting as the input parameters for machine learning models or algorithms. A feature can be described as a bit of information or observable property that is helpful to construct NLP applications or to predict NLP applications performance. The output generated from these models and algorithms are completely based on the input features and parameters.

1. Feature selection involving identification of important words

   Feature selection is the method of selecting what we consider important in the texts as well as in a sentence and determining what can be removed. The salutation and signature lines at the beginning and the end of an email contributes no value for the task of summary generation. So, it is necessary to remove these lines from the email, which we know, should not contribute to the summary. This makes for a simpler input which a model can perform better with. This will probably involve dropping punctuation and stop words, altering words by rendering them to lower case, deciding what to do with typos or grammar features.

2. Weighing the selected feature words

   TF-IDF(Term Frequency-Inverse term frequency) is a well common weighting factor which is used in natural language processing to retrieve the important features from the emails. This technique helps to identify the important words in a text and the mportance of the word increases in the same way as the count of the word that appears in a single email or document and this is called as Term Frequency. IDF or inverse document frequency is the measure of information provided by the word in the document, i.e., it shows the occurrence of the word whether its common or rare in the emails or documents.

IDF — Log(total no of documents / no of documents with the term t in it).

*So, TF-IDF = TF * IDF*

Hence, TF-IDF helps in determining the most important features from the documents or emails with its respective weights.

3. Word Embedding the features using Word2Vec and creation of feature vector

   Word embedding is skillful of obtaining a word's meaning in a text as well as the relationship with other terms and words based on that particular context. It is also used in order to determine the semantics from the sentences and the syntactic similarity between the words. Word2vec is one of the highly common strategies to implement a two-layer neural network for learning word embedding from a sentence in which its input is a corpus of text and a set of vectors is its output. A set of word vectors generated as the output from the word embedding combines related words in that space close to each other.

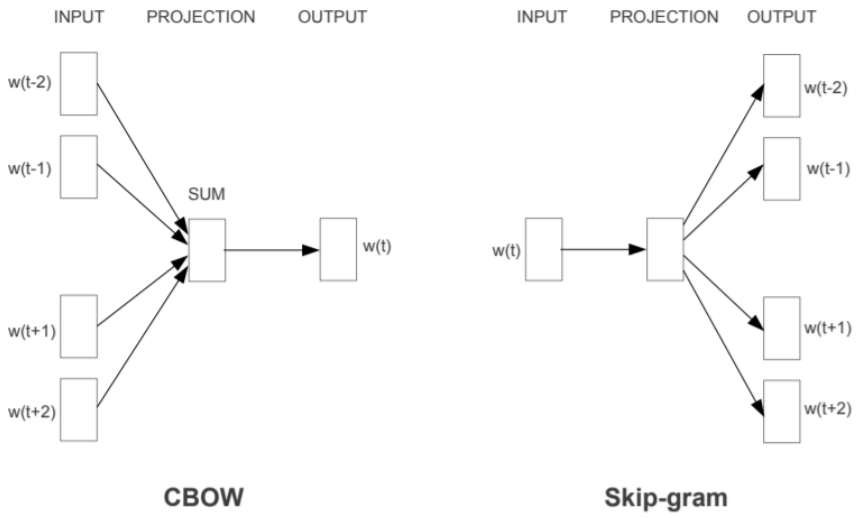


Figure 3.3: Multi-layer Word2Vec model architecture (Source:Mikolov, T., et al., 2013)

For word2vec, as it is a two-layer neural network it basically implements two algorithms for training purpose like CBOW – Continuous bag of words and skip-gram

model. There are certain significances for both the algorithms in which CBOW predicts the target word using the context and the skip-gram predicts the target context using a word. Since skip-gram method obtains two different semantics based on a single word, it generally performs much better than CBOW.

4. Word Embedding the features with BERT model and creation of feature vector

   BERT (Bidirectional Encoder Representations from Transformers) is one of the new innovations in the field of NLP which applies the bidirectional training of transformer, a common attention model for language modelling. BERT utilizes transformers by which its attention mechanism which helps to learn and identify the contextual relations among words in a text or paragraph. In the ordinary form of BERT there are two different methods that performs in transformer, initially an encoder that reads the input text and a decoder that generates the task prediction. Since the aim of BERT is to produce a language model, it only requires the encoder mechanism.

   In BERT, the input text is initially embedded into vectorized form as a series of tokens and then processed using the neural network while the output is also in vectorized sequential form where each vector resembles to an input vector of the equivalent index. There are two strategies which are used by BERT:

   (a) Masked Language Model (MLM)

   In the case of MLM technique, 15% of words in every sequence are substituted by a MASK token prior to forwarding the word sequence for proceeding with BERT and based on the non- masked words in the other context, MLM tries to predict the new value of masked words in the sequence. Technically, the prediction of new output sequence of words needs the following :

   i. Addition on top of the encoder output with a new classification layer.
   ii. The embedding matrix multiplies the output vectors and transforms them into the vocabulary component.
   iii. Calculation with softmax and the probability of every word in the dimension of vocabulary.

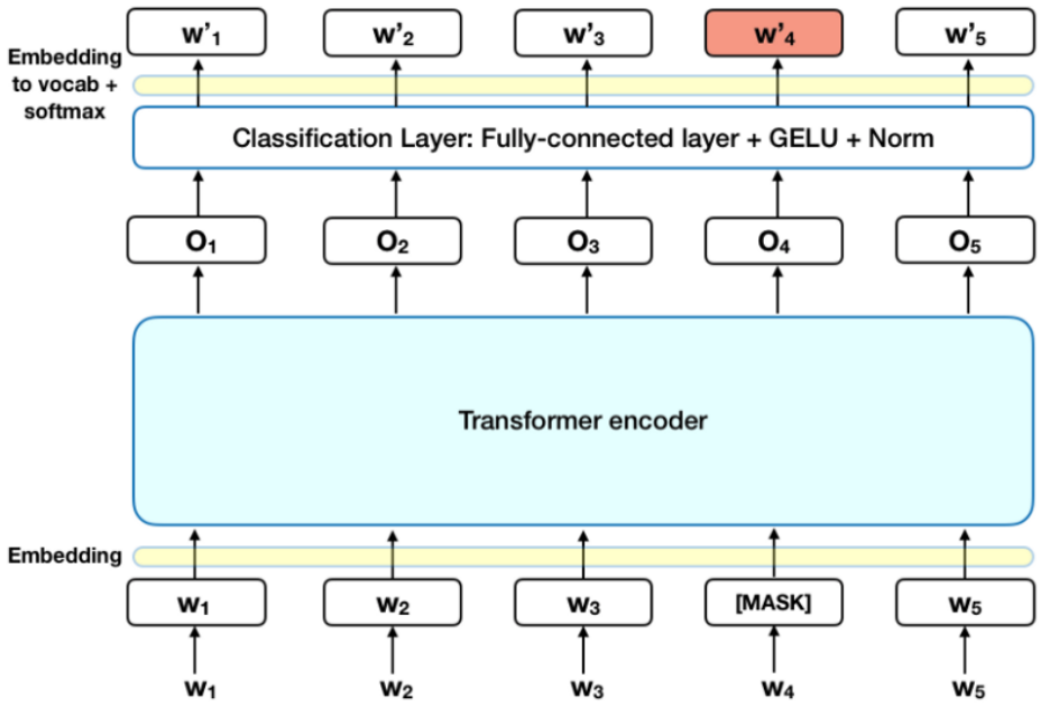


Figure 3.4: MLM Transformer model (Source: BERT [Devlin et al., 2018])

(b) Next Sentence Prediction (NSP)

During this process of training using BERT, the model obtains sentence pairs as input text and discovers to predict whether the second sentence in the pair is the following sentence of original text. In this training process, almost 50% of inputs data are in pairs in which the second sentence is the subsequent sentence in the initial corpus and as part of the later 50% random sentences from the original corpus is selected as the next sentence. It is always confirmed that the random sentence selected will always be different from the initial sentence and the possibility of recurrence is always made to null.

In order to assist the model to differentiate between the sentences and to avoid the recurrence of sentences for training set of data, the input data is managed as follows:

i. At the starting of every initial sentence, a [CLS] token is inserted and at the end of each sentence, a [SEP] token is inserted.

ii. Each token is supplemented by a sentence embedding indicating sentence

A or sentence B. Theoretically, sentence embedding is identical to token embedding with a value of 2 in vocabulary.

iii. To indicate the position of a token in the sequence, a positional embedding is included to each token.

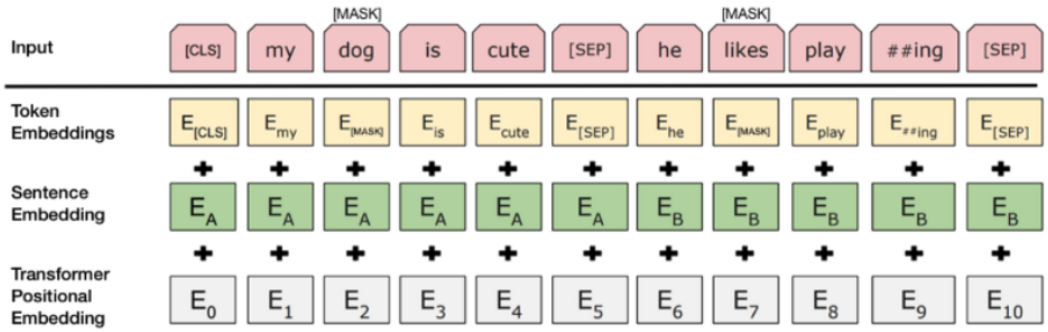


Figure 3.5: NSP Transformer model (Source: BERT [Devlin et al., 2018])

The following steps are completed to predict whether the second sentence is actually connected to the first sentence :

(a) The complete sequence of inputs data is executed through the Transformer model.

(b) A plain classification layer which are learned matrices of weights and biases are used to convert the output of [CLS] token into a vector.

(c) The probability of (IsNextSequence) is typically calculated with softmax.

The combined loss function of both the above strategies are minimized by simultaneously training Masked Language Model (MLM) and Next Sentence Prediction (NSP) together during the training process of the BERT model.

Advantages of using BERT technique for word embedding the features as compared to other traditional word embedding techniques :

(a) Traditional word embedding models used for language learning in NLP take the prior n tokens and estimate the next one. In comparison, when predicting, BERT model trains a language model that considers both the previous and next tokens.

(b) This latest BERT model is also used on later sentence prediction tasks to manage those tasks that need analysis for the connection between two sentences.

(c) BERT utilizes the Transformer design which are used certainly for encoding sentences.

(d) BERT works better when provided additional parameters, even for those datasets which are small.

3.5 Modelling

In this phase, the pre-processed data is utilized in order to implement various machine learning algorithms depending on the desired functionality. The main aim of this research is Text summarization, in which various approaches of Clustering algorithms like PHA-ClusteringGain-K-Means hybrid clustering approach and conventional K-Means algorithms are used. This Clustering phase is the next phase of the design, in which PHA clustering method is extended to PHAClusteringGain-K-Means hybrid clustering approach for efficient and accurate clustering as well as text summarization. Clustering manages the emails that have been retrieved into various groups of topics. As the terms given in the query input, the subject can be defined. Clustering strengthens the information retrieval system by providing information that is coordinated and centered. The retrieved emails are clustered during the clustering process based on their cosine-similarity scores determined in the initial stage. The method proposed will be described as follows:

3.5.1 PHA-ClusteringGain-K-Means Clustering (Hybrid approach)

In this hybrid clustering approach, Potential based Hierarchical Agglomerative clustering (PHA) algorithm and k-means clustering algorithm are merged. An evaluation benchmark criterion for clustering called clustering gain is integrated into PHA for determining the optimum count of clusters automatically. At each iterative step for PHA clustering, the clustering gain is computed, and the iteration is stopped when the clustering gain value reaches the maximum value and the number of clusters at that instant will be considered

as the ideal number of clusters for the K-means algorithm. The PHA clustering method is influenced by an issue named as chaining problems where the data objects are segregated consistently for K-means algorithm. Normally, the clusters produced from k-means are spherical/symmetrical shaped and hence the final obtained clusters form this hybrid method of clustering will be exempt from this disadvantage of chaining effect.

3.5.2 Potential-based Hierarchical Agglomerative Clustering (PHA)

Potential-based hierarchical agglomerative clustering is type of a hierarchical agglomerative clustering method by which all the data objects in the Euclidean space are accumulated in the form of clusters on the basis of hypothetical potential field among the data objects. Conventionally, HAC algorithms like Single linkage and Complete linkage clustering considers localized data information while PHA method considers both the global data information by integrating potential field and local data distribution by involving the distance matrix for the algorithm.

For PHA method, the potential between a two data objects, i, j is defined as :

$$\phi_{ij}(r_{ij}) = \begin{cases} -\frac{1}{r_{ij}} & \text{if } r_{ij} >= \delta \\ -\frac{1}{\delta} & \text{if } r_{ij} < \delta \end{cases}$$

in which r_{ij} is defined as the euclidean distance between data objects i and j and the δ parameter is determined in order to avoid the measure of singularity when r_{ij} tends to zero. The δ value can be computed as,

$$\delta = mean(MinD_i)/S$$

$$MinD_i = min_{r_{ij} \neq 0, j=1...N}(r_{ij})$$

in which $MinDi$ is considered as the minimum distance from data object i to all the other data objects which are in the same euclidean space, and S is determined as a scaling factor which is assumed as 10.

Then the total potential field at a particular data object i is calculated as the summation of potential field from all other objects till i,

$$\phi_i = \sum_{j=1...N} \phi_{ij}(r_{ij})$$

N is the count of the data objects in the euclidean space.

The step wise execution of PHA clustering is as below:

1. The data objects in euclidean space be $p1, p2, p3, p4, p5$ and $p6$, the representation of respective cosine similarity scores of the retrieved emails is displayed in the below figure.

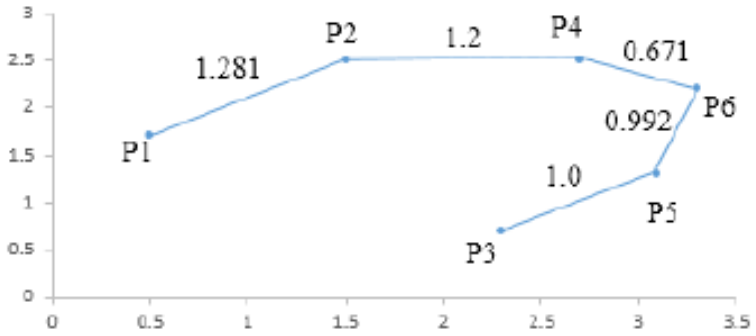


Figure 3.6: Representation of Data objects in the Euclidean space

2. The total potential at each object is calculated with the help of equation and these values of each object is arranged in ascending order.

3. After the calculation of the total potential at each data objects, let the order of potential in ascending be $[p4, p6, p5, p2, p3, p1]$, it shows that the data object $p4$ has the minimum potential and it will be the root of the Weighted Edge Tree. Similarly, this method is used to determine the parent of each object and the parent of a given object becomes the closest visited data object.

4. Since the $p4$ is the nearest visited object, the parent of $p6$ which is the second object will be set as $p4$.

5. $p6$ is considered as the parent of the third object $p5$ since the $p5$ is much closer to $p6$ than $p4$.

6. Similarly, the parent of all data objects is determined up to $p1$.

7. As per the above calculations, Euclidean distance between each data objects is determined using the weight of edge between a data object and its parent. The below diagram shows the Weighted Edge Tree.

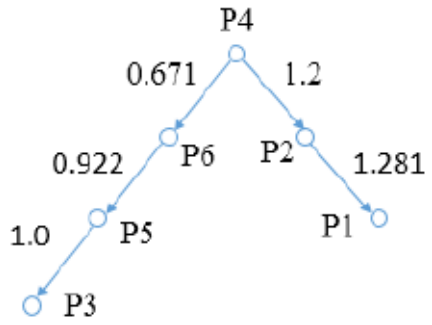


Figure 3.7: Weighted Edge Tree of data objects

8. Based on the Weighted Edge Tree, let the sorted order of data objects be $[p6, p5, p3, p2, p1, p4]$.

9. Similarly, the first data object $p6$ is merged with its parent to create a cluster of $(p6, p4)$, then second data object $p5$ is joined with $(p6, p4)$ and hence all other objects in the tree gets merged to the respective parent.

10. All the data objects are then hierarchically clustered as per every iteration and the clustering method is described with the help of dendrogram as per the below figure.

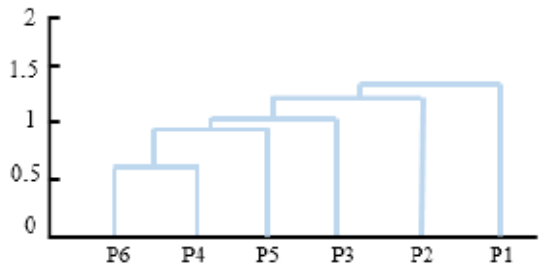


Figure 3.8: Dendrogram explaining the hierarchical clustering

Ideally, PHA clustering method has a time complexity of $O(n2)$ and hence proved that it is much effective than the traditional HAC algorithms like Single Linkage clustering which has a time complexity of $O(n3)$.

Clustering Gain

Clustering Gain is studied as one of the effective evaluation criterion parameters for determining the ideal cluster configurations and hence identifying the optimal count of clusters needed to be used for the clustering process. The clustering gain value is always higher than or equal to zero and hence the optimal number of clusters are calculated when the value of clustering gain reaches the maximum. As the clustering gain value varies at every iterations of the process, complete clustering process has to finished so as to determine the maximum clustering gain value.

At every iterations of the clustering method, Clustering gain is noted and the PHA method is converged at the highest clustering gain value. The value of the iteration when the clustering gain reaches the maximum at that stage is considered as the 'k' value for processing K-means algorithm.

During a certain stage of a clustering method in a cluster named C_j, the clustering gain Δ_j can be explained as the difference among the reduced inter-cluster error sum γ_j associated with the preceding stage and increased intra-cluster error sum λ_j associated with the prior stage.

$$\Delta_j = \gamma_j - \lambda_j$$

in which γ_j is the reduced error sum of inter-cluster as well as λ_j is the increased error sum of intra-cluster during the clustering process for the cluster named Cj.

The intra-cluster error sum named Λ can be defined as the summation of distances of all clusters, in which the distance is the summation of squared Euclidean distances to each data objects from the center of a cluster.This is also termed as within-group error sum i.e,

$$\Lambda = \sum_{j=1}^{k} \sum_{i=1}^{n_j} ||p_i^{(j)} - p_0^{(j)}||_2^2$$

in which n_j tends to the count of instances in the cluster j, k designates the number of clusters, $p_i^{(j)}$ refers to the i^{th} instance in cluster j and $p_0^{(j)}$ refers to centroidal point of cluster. The error sum value for intra-cluster increases throughout the clustering process from zero to maximum as it proceeds from the initial stage till final stage.

The error sum value of inter-cluster is named as Γ which is explained as the summation of squared Euclidean distances to the global centroid for all clusters from the center of a cluster. i.e,

$$\Gamma = \sum_{j=1}^{k} ||p_0^{(j)} - p_0||_2^2$$

p_0 is the centroid of all data objects in the cluster which is named as global centroid, i.e,

$$p_0 = \frac{1}{n} \sum_{i=1}^{n} p_i$$

3.5.3 K-Means clustering method

The k-means clustering method is processed with the calculated k value as input from the PHA algorithm. This leads to the formation of the k number of identical clusters and these output clusters created from this hybrid model will be exempt from the chaining problem.

The time complexity of this hybrid approach is O(n2) + O(nidk), i.e., the sum of time taken by both PHA and k-means methods, in which the n depicts the no. of instance in d dimension, i tells the count of instances and k is the number of clusters to be attained. If we compare with the Single linkage clustering algorithms which has time complexity O(n3), it is well efficient to use the hybrid approach.

The algorithm is composed of the following steps:

1. Place data objects as k points into the space which are being clustered, and these points signify the primary group and its respective centroids.

2. Allocate each data objects into those respective groups with the closest centroids.

3. After allocating the objects to the respective groups, the position of k centroids is recalculated.

4. The above steps are repeated until the centroids does not move or change any further. This results in the split of objects into groups from which the metric to be reduced are analyzed.

3.6 Evaluation

Validation using ROUGE metrics as the statisticalmethod : ROUGE, or Recall-Oriented Understudy for Gisting Evaluation, is a set of metrics and a software package used for evaluating automatic summarization and machine translation software in natural language processing. The metrics compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation. The measures count the number of overlapping units such as n-gram, word sequences, and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans. These above steps can be implemented using Python and Scikit-learn. We have wide varities of packages and inbuild functions as well in order to carry out the proper analysis of data using Scikit-learn.

In the case of ROUGE, the recall shows how much of the content in relation to the reference summary is retrieved or captured by the automatically summarization algorithms.A

summary (process summary) generated by a machine can be extremely long, capturing all words in the summary of the reference. Nevertheless, a lot of the words can be redundant in the system summary, rendering the summary overly verbose. In such situation Precision becomes vital. Precision depicts, how much of the information in the summary generated from algorithm are actually relevant or necessary.

Parameters	**Description**
Recall	$\frac{\textit{number of overlapping words}}{\textit{total words in the reference summary}}$
Precision	$\frac{\textit{number of overlapping words}}{\textit{total words in the automatic algorithm summary}}$

Table 3.2: Recall and Precision in context to ROUGE

The precision value enhances to be critical when we are aiming to generate summaries which are concise in nature. Thus, it is safest to calculate both the Precision value as well as the Recall values and then report the F-Measure as characteristic. If the summaries are meant to be concise due to certain constraints, then we might consider using the Recall value, as the precision value in this scenario is of less concern.

3.6.1 ROUGE-N, ROUGE-S ROUGE-L

ROUGE-N, ROUGE-S and ROUGE-L can be considered as the different levels of quality conditions for comparison of texts between automatically generated algorithm summaries and the reference summaries. For instance, ROUGE-1 implies to overlap of unigrams between the algorithm summary and reference summary while ROUGE-2 denotes to the overlap of bigrams between the reference and algorithm summaries.

- ROUGE-N – measures unigram, bigram, trigram and higher order n-gram overlap.
- ROUGE-L – refers to the longest matching sequence of words using LCS and the benefit of processing LCS is that it does not necessarily need consecutive matches,

however it matches the sequence that indicate the order of words in sentence level.

- ROUGE-S – which is also named as skip-gram cooccurrence check whether is there any pair of words amongst the sentence in order to create arbitrary gaps. For an instance, ROUGE-S measures the overlap of paired words that can have a maximum gaps of two between words in a sentence.

3.7 Strengths and Limitations

Chaining is a common problem in Single Linkage clustering and PHA method and itcan be defined as the gradual growth of a single cluster as one data object with the elementsadded to that cluster at each iterative step of the algorithm. This leads to the formation ofimpractical heterogeneous clusters and may result in the unequal partitioning of data objects.In a clustering process many singleton clusters are formed as a result of chaining. Thus, theoutput clusters cannot be properly defined from the input data objects.

The major limitation of the research is deciding optimum number of clusters k and it isthe major problem with Partitioning-based clustering approaches such as k-means method.But the proposed hybrid approach helps to overcome this limitation. Another limitation toget added up is regarding the data. There can be some email records with limited number ofwords or sentences, hence as part of implementing text summarization we have used onlythose emails records with more than 300 words which stays as a threshold for the minimumnumber of words being used for the research.

Another main limitation is rather challenging to get a public email data for researchesand studies, due to privacy related issues and GDPR establishment. This is a drawback and alimitation especially for research since the studies cannot be conducted due to unavailabilityof public email datasets. An exemption to the beyond challenge is the Enron Corpus [Klimt,Yang, (2004)], in which this email data was made public after a legal investigation regardingthe Enron Corporation.

Chapter 4

Implementation and results

This section illustrates how the research is being conducted, the different steps of the CRISP – DM methodology included in this study, the experiments conducted in the research, and the experiment results that obtained as part this study. A broad analysis of email data is performed with the help of different analytical and visual functionalities using various libraries of Python in this research.

4.1 Data Understanding

In order to plan and carry out a machine learning study, a deep understanding of the data is needed. A detailed analysis of the email data is done in this part by using various libraries of Python. The dataset used for this study is an email dataset named "Enron email data" which was publicly available for the research purpose. The Enron email dataset provides real-world text data that is arguably of the same kind as data from Echelon intercepts – a set of messages about a wide range of topics, from a large group of people. The Enron email dataset contains electronic mails in text format generated by employees of the Enron Corporation. It was obtained by the Federal Energy Regulatory Commission during its investigation of Enron's collapse.

As part of analysis for data understanding , various operations are implemented which are applicable to the text data in NLP.

4.1.1 Analytical Observations from the Data

Total number of emails:

The data used in this study contains a total of 517,401 observations with 2 features/variables both in text type data. The Enron email dataset contains electronic mails in text format generated by employees of the Enron Corporation. Since this dataset contains enormous number of email records, only limited observations are used for training and the modeling purpose and its been discussed in the further sections.

Duplicate emails:

During the research study using Unstructured data, its much important to make sure there is no data discrepancy and bias in the data. Hence, its key to find whether there is any recurrence of data and thereby determining occurrence of duplicate emails are key during analysis phase.This will help us to avoid the data redundancy as well as output models from getting biased.

Emails with no data:

In the analytical studies as we know that null values have high importance in the data since they can lead to models with discrepancy and less quality. It is not unusual for an instance or observation to be missing one or more values. Either, Information is not collected, or the attributes may not be applicable to the cases. Similarly, its key to identify those observations with no data and necessary actions can be taken so as to avoid the consequences.

Observations	Count
Total number of emails	517,401
Duplicate number of emails	0
Emails with no data	0

Table 4.1: Initial Insights from the data

Details of features/variables from the data:

The dataset using in this study is an unstructured type of textual data hence all the variables in the data is in text format. The description of initial five records in the data is being shown in the below figure:

	file	message
0	allen-p/_sent_mail/1.	Message-ID: <18782981.1075855378110.JavaMail.e...
1	allen-p/_sent_mail/10.	Message-ID: <15464986.1075855378456.JavaMail.e...
2	allen-p/_sent_mail/100.	Message-ID: <24216240.1075855687451.JavaMail.e...
3	allen-p/_sent_mail/1000.	Message-ID: <13505866.1075863688222.JavaMail.e...
4	allen-p/_sent_mail/1001.	Message-ID: <30922949.1075863688243.JavaMail.e...

Column 'file' containes just the filename, column 'message' containes multiple informations such as

- Message-ID
- Date
- From
- To
- Subject
- Mime-Version
- Content-Type
- Content-Transfer-Encoding
- X-From
- X-To
- X-cc
- X-bcc
- X-Folder
- X-Origin
- X-FileName
- Actual body

Figure 4.1: Variables description of the email data

As per the above figure we could notice that there are 2 variables and all in text format. The initial variable 'file' contains the file name which is unique, and the other variable 'message' contains a lot of information including the actual body content of the email. As for the text summarization purpose only the body content is required, and all other information had to be separated.

Sample message content in the email data is shown as below:

```
Message-ID: <13505866.1075863688222.JavaMail.evans@thyme>
Date: Mon, 23 Oct 2000 06:13:00 -0700 (PDT)
From: phillip.allen@enron.com
To: randall.gay@enron.com
Subject:
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Phillip K Allen
X-To: Randall L Gay
X-cc:
X-bcc:
X-Folder: \Phillip_Allen_Dec2000\Notes Folders\'sent mail
X-Origin: Allen-P
X-FileName: pallen.nsf

Randy,

 Can you send me a schedule of the salary and level of everyone in the
scheduling group.  Plus your thoughts on any changes that need to be made.
(Patti S for example)

Phillip
```

Figure 4.2: Sample Email format as per the data

Splitting the message body from the content:

Text summarization is a procedure for creating a summary from a text content which is the form of sentences or paragraphs. As we see in the above figure the message from the email contains surplus amount of information and hence the actual information needed for the purpose of text summarization had to be separated. A new column named 'body' was added into the data which contains only the body content which can be used for summarization purpose.

	file	message	body
0	allen-p/_sent_mail/1.	Message-ID: <18782981.1075855378110.JavaMail.e...	Here is our forecast
1	allen-p/_sent_mail/10.	Message-ID: <15464986.1075855378456.JavaMail.e...	Traveling to have a business meeting takes the...
2	allen-p/_sent_mail/100.	Message-ID: <24216240.1075855687451.JavaMail.e...	test successful. way to go!!!
3	allen-p/_sent_mail/1000.	Message-ID: <13505866.1075863688222.JavaMail.e...	Randy, Can you send me a schedule of the sal...
4	allen-p/_sent_mail/1001.	Message-ID: <30922949.1075863688243.JavaMail.e...	Let's shoot for Tuesday at 11:45.

Figure 4.3: Illustration of actual message body splitted from the data

Checking those message body with no 'X-FileName':

As we go through the email content in the data, it can be identified that there is an information named 'X-FileName' in those emails which contain a message body content and hence its mandatory to check whether any message is not having 'X-FileName' since that is our splitting benchmark. It was observed that there are no emails without 'X-FileName' information and hence it was made sure that all the observations contained the message body content.

Checking the number of forwarded information:

It is well common that emails getting forwarded from one person to another person and there will be another original email from the sender as well in the dataset. If we notice the emails from this Enron email dataset, it can be observed that the forwarded email messages contain an extra information named 'Forwarded by' in the body of the email. In order to avoid the data redundancy and avoid the repetitive information, those 'forwarded messages' had to be separated from the original 'messages' and it can used for any other analysis separately.

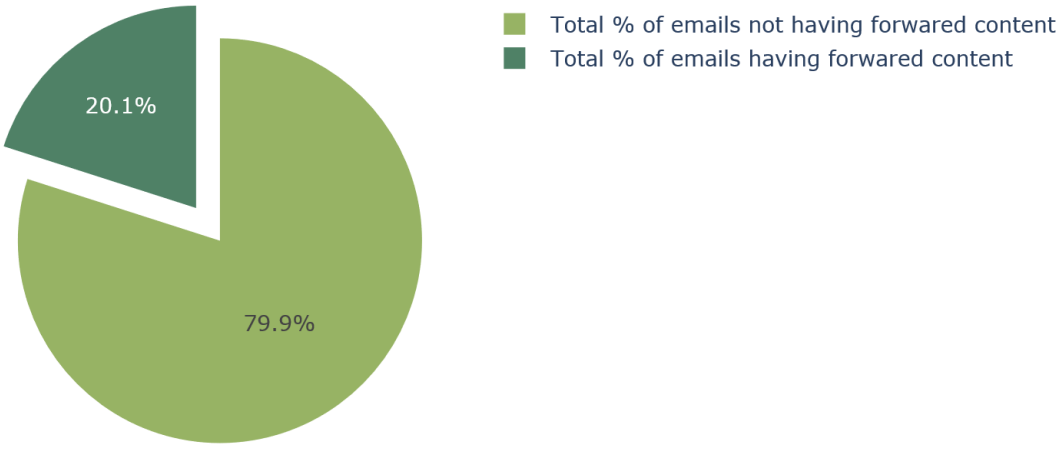


Figure 4.4: Distribution of emails in percentage based on the forwarded content

The above pie chart distribution clearly signifies that 20.1% of emails contain the forwarded contents and the rest of 79.9% data is clean with no recurrent content.

Observations	Count
Total number of emails having forwarded content	103826
Total number of emails not having forwarded content	413575

Table 4.2: Count of forwarded information from the data

As per the above observations, 103826 forwarded emails had to be processed separately since, it may contain repetitive information such as To, From, cc, bcc, body content etc.

Counting the number of words in message body contents:

As we see from the analysis, there are 413575 email records which are clean and can be used for the training purpose. But It is always recommended to use only those sufficient data that are highly relevant to the process that are implemented. Hence, identifying the word count is an important analysis for text summarization process, where the count of words is checked in order determine those emails with sufficient word counts.

Word count Observations	Values
Total Count	413575.00
Mean	329.24
Std	1761.01
Min	1.00
25%	41.00
50%	108.00
75%	255.00
Max	280945.00

Table 4.3: Probabilistic distributions of word count from the data

From the above probabilistic observations, a benchmark word count for text summarization purpose was determined and only those email samples with a minimum word count of 329 words are considered for processing. Hence, emails with lower word count are removed and separated from the data which are used for processing and modelling.

Observations	Count
Number of samples with word count greater than mean word count (329) value	79565

Table 4.4: Determining the samples which can be used for the processing

4.2 Data Preparation

As part of Data Understanding in the above section, we have determined the sample which are applicable for further processing and analysis purpose.With reference to Table 4.4, 79565 number of samples with word count greater than mean word count (329) value are used in this data preparation procedure and henceforth. As per the analysis of the email contents, the data should undergo some cleaning and preparation steps like :

4.2.1 Spelling Check and correction

As part of initial cleaning purpose, spell check is performed in which each and every word in the email contents are analyzed and their respective spellings are checked. From the email content, all the words in the range '$[\wedge A - Za - z]'$ are separated and the corresponding spell check is done. This function is provided by the python package 'pyspellchecker' to discover the words that could have mis-spelled and indicate likely corrections. It leads to import library called 'SpellChecker' and the corresponding function helps to determine the wrongly spelled words and provides some suggestions regarding possible appropriate words. Another function used are 'spell.correction' which helps to correct the word using SpellChecker library and it returns the most likely result for the words that are wrongly spelled.

4.2.2 Named entity recognition (NER)

There are specific terms in any text document or in any text paragraphs that signify particular entities which are used with respect to a unique context and are much informative. These informative entities are termed as named entities which more precisely indicate those real world things such as places, peoples, organizations, and so on, and these are represented with unique names, hence it is vital to recognize these informative entities. A prevalent approach used for information retrieval in order to determine the named entities as well as to segregate them to categorize these entities into different classes which are predefined. This technique is called as Named entity recognition (NER).

Various python libraries support this technique and a library named as 'Spacy' has several exceptional abilities for named entity recognition. This procedure is implemented along with the spell checker functionality as discussed in the above, in which the entire text email message is used to identify the mis-spelled words and the NER procedure is performed on all these mis-spelled words so that various important entities like names, places, organization names will not treated as mis-spelled words.

4.2.3 Abbreviations correction

Abbreviations/acronyms are always a part of written as well as text communication irrespective of the domain and moreover, in sentences that use them, abbreviations are not generally specified. Identifying and understanding the abbreviations that are used in a particular sentence frequently needs broad knowledge with respect to the target domain and the capability to recognize them context-based. Using abbreviations and acronyms can cause difficulty for people that are new in a particular field to read and understand and also it can lead to confusion and challenging in the field of text processing.

In order to handle this problem, we use the abbreviation-expansion content that are collected [1], in the format of .xlsx and are incorporated in order to replace the abbreviations/acronyms in the email body contents. The abbreviation and the respective expansions from the file are combined into a dictionary using python and then each word from the

[1] https://www.webopedia.com/quick_ref/textmessageabbreviations.asp

emails body are scanned, the abbreviations are identified and hence those abbreviations are replaced with the expansions from the dictionary.

abbrevation	expansion
afaiui	as far as i understand it
afap	as far as possible
affa	angels forever, forever angels
afj	april fool's joke
afk	away from keyboard

Figure 4.5: Initial 5 Sample of the abbreviation-expansions

4.2.4 Case Lowering, Remove Punctuations and Tokenization

From a paragraph of texts there are some chances for the mis use of lower case and upper case, in this all the sentences are being reformed into lower cases and all the punctuations are removed. NLTK Tokenizer package in Python helps to divide strings into lists of substrings. It can be used to find the words and punctuation in a string and this is basically the initial step to be performed in NLP as part of prep-rocessing. This process consists of three phases like dividing the sentences into words, realizing the significance of each word with reverence to the respective sentence and ultimately generate structural explanation on the input sentence.

4.2.5 Stemming/Lemmatization

Stemming and Lemmatization are Text Normalization process or referred to as Word Normalization methods in the area of NLP which are used primarily to organize texts, words and sentences in documents for advanced processing. Stemming algorithms are empirical methods work by cutting the end or beginning of a word in the hope of the goal correctly, considering a list of mutual prefixes and suffixes that can be contained in an inflected term and words. Porter's stemming algorithm is a well common algorithm which is used to perform stemming of words from the sentence. Stemming can reduce the memory space required to store the words and makes computation easier. While Lemmatization, on the other

hand, considers the morphological study of words by which it is important to have comprehensive dictionaries that the algorithm can consider through to connect the form back to its lemma factor.

4.3 Data Modelling

Once the data preparation and the data cleaning are completed, the same can be used for further modelling stage. This Modelling stage comprises of generating the feature vectors using different word embedding techniques along with applying the clustering algorithms in the embedded features. The data available after both the above phases are completely ready to be used for implementing word embedding techniques for generating the feature vectors and for applying clustering algorithms for generating the text summaries of the email body contents. Hence the cleaned data has been written to an excel file and are further used for the modelling and experimentation purpose. This excel file used for training and testing purposes consist of various elements like cleaned/pre-processed email content, number of clusters and the manual summary/Reference summary.

email	number of clusters	manual summary
"Liane, As we discussed yesterday, I am conc...	5	At the time of these trades, offers for physic...
George, Below is a list of questions that Ke...	8	George, Below is a list of questions that Keit...
"Reagan, Thank you for the quick response on ...	9	What type of floor joist would be used?\r\n\r\...
"George, I am back in the office and ready to...	7	Specifically that the costs of our project are...
Lucy, Please fix #41 balance by deleting the ...	5	The other questions I had about last week\s re...

Figure 4.6: Sample data format used for modelling process

As per the above figure, the email contents in the file is the pre-processed and cleaned emails and the manual summary is the reference summary which is generated using an online tool[2]. Both these summaries are used for the evaluation purpose using ROUGE method. The number of clusters that are assigned in the file is dependent on the clustering algorithm that are being implemented. The further details of clustering are explained in the below experiments.

[2]http://autosummarizer.com/index.php

There are different approaches being implemented in order to answer the research question and satisfy the aim of the research. The various experiments carried out in this data modelling phase are detailed below:

4.3.1 Experiment 1 : Word2Vec + K-means clustering

In this experiment word2vec model was used as the word embedding for feature engineering and it was incorporated to K-means clustering for generating the text summaries. As part of tokenizing, TF-IDF vectorizer was used and then these word tokens were being passed on to genism model for creating the feature vectors for training purpose. Separate functions were written to generate the feature embedding, implement the text summarization and the evaluation purpose using ROUGE method.

As part of text summarization purpose the cleaned email content from the file was extracted and the paragraph content from each mail was split line by line and all the words were made into lower case before processed for clustering. Few parameters were set for applying the clustering algorithm, in the case of k- means clustering, the value for 'k' i.e, the number of clusters were set based on the length of contents in the email body.We can also determine the k value by setting the value for percentage of compression. Once these parameters are set, then the lines from the paragraphs of email are passed for word embedding using word2vec model. The process of word embedding is followed after the weighing of the selected feature words using TF-IDF vectorizer. This process of word embedding using word2vec returns the sentence vectors. Along with the number of clusters set as the parameters and the sentence vectors obtained from the word2vec feature embedding, the K-means clustering algorithm is applied. Based on the closest centroid and the minimum pairwise distances between the sentence vectors, the summary was generated.

The following Table 4.5 illustrates the Original email content from Enron email dataset, Reference summary obtained using online tool and the Model summary generated using the K- means clustering with Word2vec as feature embedding method of a sample email.

Type	Content
Original Email	”George, The probability of building a house this year is increasing. I have shifted to a slightly different plan. There were too many design items that I could not work out in the plan we discussed previously. Now, I am leaning more towards a plan with two wings and a covered courtyard in the center. One wing would have a living/dining kitchen plus master bedroom downstairs with 3 kid bedrooms + a laundry room upstairs. The other wing would have a garage + guestroom downstairs with a game room + office/exercise room upstairs. This plan still has the same number of rooms as the other plan but with the courtyard and pool in the center this plan should promote more outdoor living. I am planning to orient the house so that the garage faces the west. The center courtyard would be covered with a metal roof with some fiberglass skylights supported by metal posts. I am envisioning the two wings to have single slope roofs that are not connected to the center building. I don’t know if you can imagine the house I am trying to describe. I would like to come and visit you again this month. If it would work for you, I would like to drive up on Sunday afternoon on Feb. 18 around 2 or 3 pm. I would like to see the progress on the house we looked at and tour the one we didn’t have time for. I can bring more detailed drawings of my new plan. Call or email to let me know if this would work for you. pallen70@hotmail.com or 713-463-8626(home), 713-853-7041(work) Phillip Allen PS. Channel 2 in Houston ran a story yesterday (Feb. 2) about a home in Kingwood that had a poisonous strain of mold growing in the walls. You should try their website or call the station to get the full story. It would makes a good case for breathable walls.”
Reference Summary	George, The probability of building a house this year is increasing. I have shifted to a slightly different plan. Now, I am leaning more towards a plan with two wings and a covered courtyard in the center. This plan still has the same number of rooms as the other plan but with the courtyard and pool in the center this plan should promote more outdoor living. I am planning to orient the house so that the garage faces the west. I am envisioning the two wings to have single slope roofs that are not connected to the center building. I dont know if you can imagine the house I am trying to describe.
Model Summary	George, This plan still has the same number of rooms as the other plan but with the courtyard and pool in the center this plan should promote more outdoor living. I don’t know if you can imagine the house i am trying to describe. I would like to come and visit you again this month on feb. 18 around 2 or 3 pm. Call or email to let me know if this would work for you. It would makes a good case for breathable walls.

Table 4.5: Sample Model Summary: Word2Vec + K-means clustering

Result of Experiment 1:

In order to obtain the final results, the summary generated from the k-means clustering and the reference summaries generated as discussed above are being evaluated and compared. This is achieved using ROUGE-N technique, in which the summary generated from the clustering algorithm and the reference summary generated from the online tool was compared and evaluated. A bi-gram ROUGE technique was executed with N value as 2, which refers to refers to the overlap of bigrams among the algorithm generated summaries and reference summaries.

Measures	Values
Recall	0.3706
Precision	0.4750
F1-measure	0.3640

Table 4.6: ROUGE-2 results of Experiment 1 - Word2Vec + K-means clustering

Once the ROUGE-2 technique is computed between the algorithm generated summaries and reference summaries, we get the individual Recall, Precision and F1- measure of respective emails. The above Table 4.6 displays the Average/Mean values of Recall, Precision and F1-measure of all emails which was used as test data.

4.3.2 Experiment 2 : Word2Vec + PHA-ClusteringGain-K-means clustering

This approach of experiment also used word2vec model as the word embedding for feature engineering and it was fused to a Hybrid clustering model using PHA-ClusteringGain-K-means clustering for generating the text summaries. As part of tokenizing, TFIDF-vectorizer was used and then these word tokens were being passed on to genism model for creating the feature vectors for training purpose. Separate functions were written to implement the text summarization and the evaluation purpose using ROUGE method.

For the text summarization purpose the cleaned email content from the file was extracted and the paragraph content from each mail was split line by line and all the words were made into lower case before processed for clustering. The lines from the paragraphs of email are passed for word embedding using word2vec model. The process of word embedding is followed after the weighing of the selected feature words using TF-IDF vectorizer. This process of word embedding using word2vec returns the sentence vectors. In the case of PHA-ClusteringGain-K-means clustering, PHA agglomerative clustering was used for determining the clustering gain, once the value of clustering gain reaches the maximum value over the iterations then the number of clusters will be fixed. This fixed number of clusters at the maximum clustering gain was given as the value for k in the k-means algorithm. Once the k value is fixed, then the k-means clustering is performed using the sentence embedding generated using the word2vec model and the number of clusters set using the PHA-ClusteringGain algorithm. Based on the closest centroid and the minimum pairwise distances between the sentence vectors, the summary was generated.

Result of Experiment 2:

In order to obtain the final results, the summary generated from the PHA-ClusteringGain-K-means clustering and the reference summaries generated as discussed above are being evaluated and compared. This is achieved using ROUGE-N technique, in which the summary generated from the clustering algorithm and the reference summary generated from the online tool was compared and evaluated. A bi-gram ROUGE technique was executed with N value as 2, which refers to refers to the overlap of bigrams among the algorithm generated summaries and reference summaries.

Measures	Values
Recall	0.2725
Precision	0.5573
F1-measure	0.3213

Table 4.7: ROUGE-2 results of Experiment 2 - Word2Vec + (PHA-ClusteringGain-K-means)

Type	Content
Original Email	”George, The probability of building a house this year is increasing. I have shifted to a slightly different plan. There were too many design items that I could not work out in the plan we discussed previously. Now, I am leaning more towards a plan with two wings and a covered courtyard in the center. One wing would have a living/dining kitchen plus master bedroom downstairs with 3 kid bedrooms + a laundry room upstairs. The other wing would have a garage + guestroom downstairs with a game room + office/exercise room upstairs. This plan still has the same number of rooms as the other plan but with the courtyard and pool in the center this plan should promote more outdoor living. I am planning to orient the house so that the garage faces the west. The center courtyard would be covered with a metal roof with some fiberglass skylights supported by metal posts. I am envisioning the two wings to have single slope roofs that are not connected to the center building. I don’t know if you can imagine the house I am trying to describe. I would like to come and visit you again this month. If it would work for you, I would like to drive up on Sunday afternoon on Feb. 18 around 2 or 3 pm. I would like to see the progress on the house we looked at and tour the one we didn’t have time for. I can bring more detailed drawings of my new plan. Call or email to let me know if this would work for you. pallen70@hotmail.com or 713-463-8626(home), 713-853-7041(work) Phillip Allen PS. Channel 2 in Houston ran a story yesterday (Feb. 2) about a home in Kingwood that had a poisonous strain of mold growing in the walls. You should try their website or call the station to get the full story. It would makes a good case for breathable walls.”
Reference Summary	George, The probability of building a house this year is increasing. I have shifted to a slightly different plan. Now, I am leaning more towards a plan with two wings and a covered courtyard in the center. This plan still has the same number of rooms as the other plan but with the courtyard and pool in the center this plan should promote more outdoor living. I am planning to orient the house so that the garage faces the west. I am envisioning the two wings to have single slope roofs that are not connected to the center building. I dont know if you can imagine the house I am trying to describe.
Model Summary	This plan still has the same number of rooms as the other plan but with the courtyard and pool in the center this plan should promote more outdoor living.I would like to drive up on Sunday afternoon on Feb. 18 around 2 or 3 pm. I would like to see the progress on the house we looked at and tour the one we didn t have time for.

Table 4.8: Sample Model Summary: Word2Vec + (PHA-ClusteringGain-K-means)

The above Table 4.8 illustrates the Original email content from Enron email dataset, Reference summary obtained using online tool and the Model summary generated using the PHA-ClusteringGain-K-means clustering with Word2vec as feature embedding method of a sample email.

Once the ROUGE-2 technique is computed between the algorithm generated summaries and reference summaries, we get the individual Recall, Precision and F1- measure of respective emails. The above Table 4.7 displays the Average/Mean values of Recall, Precision and F1-measure of all emails which was used as test data for Word2Vec + PHA-ClusteringGain-K-means clustering.

4.3.3 Experiment 3 : BERT + K-means clustering

As part of different experimental approaches in this BERT model was used as the word embedding for feature engineering and it was used apply the K-means clustering for generating the text summaries. For this purpose, a python library called "bert_embedding" was used in order to implement the BERT model for creating the feature vectors for training with algorithms. Separate functions were written to generate the feature embedding, implement the text summarization and the evaluation purpose using ROUGE method.

During the clustering and text summarization phase, the cleaned email content from the file was extracted and those paragraph content from each mail was split line by line and all the words were made into lower case before processed for clustering. Few parameters were set for applying the clustering algorithm, in the case of k- means clustering, the value for 'k' i.e., the number of clusters were set based on the length of contents in the email body. We can also determine the k value by setting the value for percentage of compression. Once these parameters are set, then the lines from the paragraphs of email are passed for word embedding using BERT model. This process of word embedding using BERT model returns the sentence vectors. Along with the number of clusters set as the parameters and the sentence vectors obtained from the BERT feature embedding, the K-means clustering algorithm is applied. Based on the closest centroid and the minimum pairwise distances between the sentence vectors, the summary was generated.

Type	Content
Original Email	”George, The probability of building a house this year is increasing. I have shifted to a slightly different plan. There were too many design items that I could not work out in the plan we discussed previously. Now, I am leaning more towards a plan with two wings and a covered courtyard in the center. One wing would have a living/dining kitchen plus master bedroom downstairs with 3 kid bedrooms + a laundry room upstairs. The other wing would have a garage + guestroom downstairs with a game room + office/exercise room upstairs. This plan still has the same number of rooms as the other plan but with the courtyard and pool in the center this plan should promote more outdoor living. I am planning to orient the house so that the garage faces the west. The center courtyard would be covered with a metal roof with some fiberglass skylights supported by metal posts. I am envisioning the two wings to have single slope roofs that are not connected to the center building. I don’t know if you can imagine the house I am trying to describe. I would like to come and visit you again this month. If it would work for you, I would like to drive up on Sunday afternoon on Feb. 18 around 2 or 3 pm. I would like to see the progress on the house we looked at and tour the one we didn’t have time for. I can bring more detailed drawings of my new plan. Call or email to let me know if this would work for you. pallen70@hotmail.com or 713-463-8626(home), 713-853-7041(work) Phillip Allen PS. Channel 2 in Houston ran a story yesterday (Feb. 2) about a home in Kingwood that had a poisonous strain of mold growing in the walls. You should try their website or call the station to get the full story. It would makes a good case for breathable walls.”
Reference Summary	George, The probability of building a house this year is increasing. I have shifted to a slightly different plan. Now, I am leaning more towards a plan with two wings and a covered courtyard in the center. This plan still has the same number of rooms as the other plan but with the courtyard and pool in the center this plan should promote more outdoor living. I am planning to orient the house so that the garage faces the west. I am envisioning the two wings to have single slope roofs that are not connected to the center building. I dont know if you can imagine the house I am trying to describe.
Model Summary	George, the probability of building a house this year is increasing. Now, I am leaning more towards a plan with two wings and a covered courtyard in the center. The center courtyard would be covered with a metal roof with some fiberglass skylights supported by metal posts. I would like to come and visit you again this month Feb. 18 around 2 or 3 pm. You should try their website or call the station to get the full story.

Table 4.9: Sample Model Summary: BERT + K-means clustering

The above Table 4.9 illustrates the Original email content from Enron email dataset, Reference summary obtained using online tool and the Model summary generated using the K-means clustering with BERT model as feature embedding method of a sample email.

Result of Experiment 3:

In order to obtain the final results, the summary generated from the k-means clustering and the reference summaries generated as discussed above are being evaluated and compared. This is achieved using ROUGE-N technique, in which the summary generated from the clustering algorithm and the reference summary generated from the online tool was compared and evaluated. A bi-gram ROUGE technique was executed with N value as 2, which refers to refers to the overlap of bigrams among the algorithm generated summaries and reference summaries.

Measures	Values
Recall	0.4182
Precision	0.4714
F1-measure	0.3915

Table 4.10: ROUGE-2 results of Experiment 3 - BERT + K-means clustering

Once the ROUGE-2 technique is computed between the algorithm generated summaries and reference summaries, we get the individual Recall, Precision and F1- measure of respective emails. The above Table 4.10 displays the Average/Mean values of Recall, Precision and F1-measure of all emails which was used as test data.

4.3.4 Experiment 4 : BERT + PHA-ClusteringGain-K-means clustering

This approach of experiment was to implement the BERT model as the word embedding for feature engineering and it was fused to a Hybrid clustering model using PHA-ClusteringGain-K-means clustering for generating the text summaries. For this purpose, a python library called “bert_embedding” was used in order to implement the BERT model for creating the

feature vectors for training with algorithms and separate functions were written to implement the text summarization and the evaluation purpose using ROUGE method.

For the text summarization purpose the cleaned email content from the file was extracted and the paragraph content from each mail was split line by line and all the words were made into lower case before processed for clustering. The lines from the paragraphs of email are passed for word embedding using BERT model. This process of word embedding using BERT model returns the sentence vectors. In the case of PHA-ClusteringGain-K-means clustering, PHA agglomerative clustering was used for determining the clustering gain, once the value of clustering gain reaches the maximum value over the iterations then the number of clusters will be fixed. This fixed number of clusters at the maximum value of clustering gain was given as the value for k in the k-means algorithm. Once the k value is fixed, then the k-means clustering is performed using the sentence vectors generated using the BERT model and the number of clusters set using the PHA-ClusteringGain algorithm. Based on the closest centroid and the minimum pairwise distances between the sentence vectors, the summary was generated.

Result of Experiment 4:

In order to obtain the final results, the summary generated from the PHA-ClusteringGain-K-means clustering and the reference summaries generated as discussed above are being evaluated and compared. This is achieved using ROUGE-N technique,a bi-gram ROUGE technique was executed with N value as 2, which refers to refers to the overlap of bigrams among the algorithm generated summaries and reference summaries.

Measures	Values
Recall	0.2860
Precision	0.5431
F1-measure	0.3293

Table 4.11: ROUGE-2 results of Experiment 4 - BERT + (PHA-ClusteringGain-K-means)

Type	Content
Original Email	”George, The probability of building a house this year is increasing. I have shifted to a slightly different plan. There were too many design items that I could not work out in the plan we discussed previously. Now, I am leaning more towards a plan with two wings and a covered courtyard in the center. One wing would have a living/dining kitchen plus master bedroom downstairs with 3 kid bedrooms + a laundry room upstairs. The other wing would have a garage + guestroom downstairs with a game room + office/exercise room upstairs. This plan still has the same number of rooms as the other plan but with the courtyard and pool in the center this plan should promote more outdoor living. I am planning to orient the house so that the garage faces the west. The center courtyard would be covered with a metal roof with some fiberglass skylights supported by metal posts. I am envisioning the two wings to have single slope roofs that are not connected to the center building. I don’t know if you can imagine the house I am trying to describe. I would like to come and visit you again this month. If it would work for you, I would like to drive up on Sunday afternoon on Feb. 18 around 2 or 3 pm. I would like to see the progress on the house we looked at and tour the one we didn’t have time for. I can bring more detailed drawings of my new plan. Call or email to let me know if this would work for you. pallen70@hotmail.com or 713-463-8626(home), 713-853-7041(work) Phillip Allen PS. Channel 2 in Houston ran a story yesterday (Feb. 2) about a home in Kingwood that had a poisonous strain of mold growing in the walls. You should try their website or call the station to get the full story. It would makes a good case for breathable walls.”
Reference Summary	George, The probability of building a house this year is increasing. I have shifted to a slightly different plan. Now, I am leaning more towards a plan with two wings and a covered courtyard in the center. This plan still has the same number of rooms as the other plan but with the courtyard and pool in the center this plan should promote more outdoor living. I am planning to orient the house so that the garage faces the west. I am envisioning the two wings to have single slope roofs that are not connected to the center building. I dont know if you can imagine the house I am trying to describe.
Model Summary	Now, I am leaning more towards a plan with two wings and a covered courtyard in the center. One wing would have a living or dining kitchen plus master bedroom downstairs with kid bedrooms and a laundry room upstairs. I dont know if you can imagine the house I am trying to describe. I would like to come and visit you again this month.

Table 4.12: Sample Model Summary: BERT + (PHA-ClusteringGain-K-means)

The above Table 4.12 illustrates the Original email content from Enron email dataset, Reference summary obtained using online tool and the Model summary generated using the PHA-ClusteringGain-K-means clustering with BERT as feature embedding method of a sample email.

Once the ROUGE-2 technique is computed between the algorithm generated summaries and reference summaries, we get the individual Recall, Precision and F1- measure of respective emails. The above Table 4.11 displays the Average/Mean values of Recall, Precision and F1-measure of all emails which was used as test data for BERT + PHA-ClusteringGain-K-means clustering.

Chapter 5

Evaluation and discussion

This chapter is used to analysis of results which are obtained as part of different experimental approaches in the research within context to aim of the research and the problem definition. This deals with the comparison of various results obtained from various experiments conducted during the implementation part of Chapter 4. The Objective of this research was to conduct various combinations of feature embedding techniques such as word2vec/BERT model along with a hybrid clustering model named PHA-ClusteringGain-K-means clustering and conventional k means clustering algorithm and to determine whether the hybrid approach outperforms the conventional approach in text summarization.

5.1 Evaluation of Experiments

This section evaluates in more detail the experiments carried out in Chapter 4. A series of experiments were conducted within the context to the research problem and to determine whether the hybrid approach of clustering will outperform the conventional approach in text summarization when different combinations of feature embedding techniques are applied. The results of the experiments are analyzed in this section.

The evaluation of the experiments was conducted on the basis of results obtained using ROUGE method. This method introduced by [Lin (2004)] is also termed as Recall Oriented Understudy of Gisting Evaluation (ROUGE), it is a set of metrics which provides the scores on the basis of similarities with the sequence of words amongst a manual/online summary

and an automated/algorithmic summary. This method helps to automatically evaluate the summary and they are in five different forms such as ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S and ROUGE-SU. Throughout the experiments ROUGE-N was used with 'N' value as 2 also called as bi-gram which measures the bi-gram units shared amongst an automated summary and a reference summary.When it comes to the evaluation analysis it is all about the scores generated using the ROUGE technique and these are Precision, Recall and F1-measure.

Experiments	**Recall**	**Precision**	**F1-measure**
Word2Vec+K_means	0.3706	0.4751	0.3641
Word2Vec+PHA_Kmeans	0.2725	0.5573	0.3213
BERT+K_means	0.4182	0.4714	0.3915
BERT+PHA_Kmeans	0.2861	0.5431	0.3293

Table 5.1: Comparison of experiments based on ROUGE-2 metrics scores

The above Table 5.1 shows the respective results which are obtained as part of all the experiments which were conducted using different approaches and combinations of feature embedding techniques and clustering algorithms.

1. **Precision**: It determines what portion of the sentences chosen by the humans/reference summary and selected by the algorithmic/automated summary are correct. Precision also talks about the percentage of n-grams in the algorithmic summary which are common with the reference summary. It is mathematically calculated as the number of sentences found in both algorithmic and reference summaries divided by the number of sentences in the algorithmic summary.

2. **Recall**: It determines what amount of the sentences chosen by humans/reference is even recognized by the machine/algorithm. Recall also conveys the percentage of n-grams in the reference summary which are in common with the algorithmic generated summary. It is the number of sentences found in both algorithmic and reference

summaries divided by the number of sentences in the reference summary.

3. **F1-measure**: It is normally computed by using both recall and precision values.

The first experiment was based on the implementation of Word2Vec feature embedding followed by conventional K- means clustering with the test data. Based on the benchmark cutoff for creating test data, email contents with more than 329 words was selected and a cleaned test data was created. In this approach of using conventional k-means clustering, the k value which designates the number of clusters was set manually depending on the length and lines in the paragraphs of email body. Then the cleaned test data was prepared and was forwarded for applying Word2Vec, test data was prepared and was forwarded for applying Word2Vec feature embedding which generated the sentence vectors. The conventional k-means clustering algorithm was executed on the sentence vectors which were generated after the feature embedding. Then the summary generated by the algorithm was evaluated with the reference summary using ROUGE-2 technique. The final results obtained as part this experiment gave a Recall value of 0.3706, Precision of 0.4751 and a f1 measure of 0.3641. The recall value conveys that, 37.06% of sentences available in the reference summary was recognized by the algorithmic summary which was generated using word2vec feature embedding followed by the k-means clustering. This also shows that this much percentage of bi-grams which was in the reference summary are also present in the automatically generated algorithmic summary. If we look into the precision value obtained from this approach, it shows that 47.51% of sentences chosen in the reference summary and the algorithmic summary are correct and the same percentage of bi-grams in the generated algorithmic summary are in common with the reference summary. The computed combination value of both the recall and precision value was obtained as 36.41% in the form of F1-measure.

Experiment 2 was performed on the basis of a hybrid approach in clustering, in which an Agglomerative hierarchical clustering algorithm was merged along with a conventional k-means clustering algorithm which came to be known as “PHA-ClusteringGainK-means” clustering algorithm. The reason behind this approach was to set the ‘k’ value i.e., number

of clusters for conventional k means clustering algorithm with the help of another Hierarchical clustering algorithm instead of setting the k value manually. The cleaned test data was used in order to perform this experiment and word embedding technique named as word2vec was applied for selecting the feature embedding vectors. These sentence vectors from the test data was forwarded to the hybrid algorithm for performing clustering and generating the summary. This generated summary was evaluated against the reference summary using ROUGE-2 technique. The results obtained as part of evaluation provided Recall value of 0.2725, Precision of 0.5573 and f1-measure of 0.3213.

The recall value from this experiment using hybrid approach reveals that, 27.25% of sentences that was present in the reference summary was identified by the algorithmic summary which was generated using word2vec feature embedding followed by PHA-ClusteringGainK-means clustering algorithm. It also shows that this much percentage of bi-grams from the reference summary are also present in the summary which was generated using clustering algorithm. Even though this recall value is less, another interesting fact from the results is about the precision value. The ROUGE method shows that with this approach of experimentation, a precision of 0.5573 was obtained and it depicts that 55.73% of sentences chosen in the reference summary and the algorithmic summary are correct and it proves that this approach generates more precise summary than the conventional approach. It also shows that the same percentage of bi-grams in the generated algorithmic summary are in common with the reference summary. A computed combination of precision and recall value of 32.13%was obtained as f1-measure.

The third experiment was based on the implementation of a latest technique in the field of NLP named as BERT model for selecting feature embedding vectors followed by conventional K- means clustering with the test data. In this approach of using conventional k-means clustering, the k value which designates the number of clusters was set manually depending on the length and lines in the paragraphs. Then the cleaned test data was prepared and was forwarded for applying BERT model for feature embedding which generated the sentence vectors. The conventional k-means clustering algorithm was executed on the sentence vectors which were generated in the process of feature embedding. Then the sum-

mary generated by the conventional k-means algorithm was evaluated with the reference summary using ROUGE-2 technique. The final results obtained as part this experiment yielded a Recall value of 0.4182, Precision of 0.4714 and a f1 measure of 0.3915. The recall value conveys that, 41.82% of sentences available in the reference summary was known by the algorithmic summary which was generated by the k-means clustering using BERT technique. This also shows that this much percentage of bi-grams which was in the reference summary are also appeared in the automatically generated algorithmic summary. The results also generated a precision percentage of 47.14%, which depicts that the sentences chosen in the reference summary and the algorithmic summary are correct and the same percentage of bi-grams in the generated algorithmic summary are in common with the reference summary. The computed and combined value of both the recall and precision through f1-measure was found to be 39.15%.

The final experiment was performed also on the basis of a hybrid approach in clustering, in which an Agglomerative hierarchical clustering algorithm was combined along with a conventional k-means clustering algorithm which came to be known as "PHA-ClusteringGainK-means" clustering algorithm. The reason behind this approach was to set the 'k' value i.e., number of clusters for conventional k means clustering algorithm with the help of another Hierarchical clustering algorithm instead of setting the k value manually. The difference between this experiment and the experiment-2 was with the technique used in for feature embedding. In this experiment, BERT model was used to implement the process of feature embedding and for selecting sentence vectors from the test data for clustering. These sentence vectors from the test data was forwarded to the hybrid algorithm for performing clustering and generating the summary. This generated summary was evaluated against the reference summary using ROUGE-2 technique. The results obtained as part of evaluation provided Recall value of 0.2861, Precision of 0.5431 and f1-measure of 0.3293.

The recall value from this experiment using hybrid approach reveals that, 28.61% of sentences that was present in the reference summary was detected by the algorithmic summary which was generated using BERT model for feature embedding followed by PHA-ClusteringGainK-means clustering algorithm. It also indicates that this much percentage of

bi-grams from the reference summary are also present in the summary which was generated using clustering algorithm. Even though this recall is lightly acceptable, another interesting fact again from the results is about the precision value. The ROUGE method shows that with this approach of experimentation, a precision of 0.5431 was obtained and it depicts that 54.31% of sentences chosen in the reference summary and the algorithmic summary are correct and it proves that this approach generates more precise summary than the conventional approach. It also reveals that the same percentage of bi-grams in the generated algorithmic summary are in common with the reference summary. A computed combination of precision and recall value of 32.93%was obtained as f1-measure from this experimental approach.

5.2 Individual Evaluation of ROUGE measures

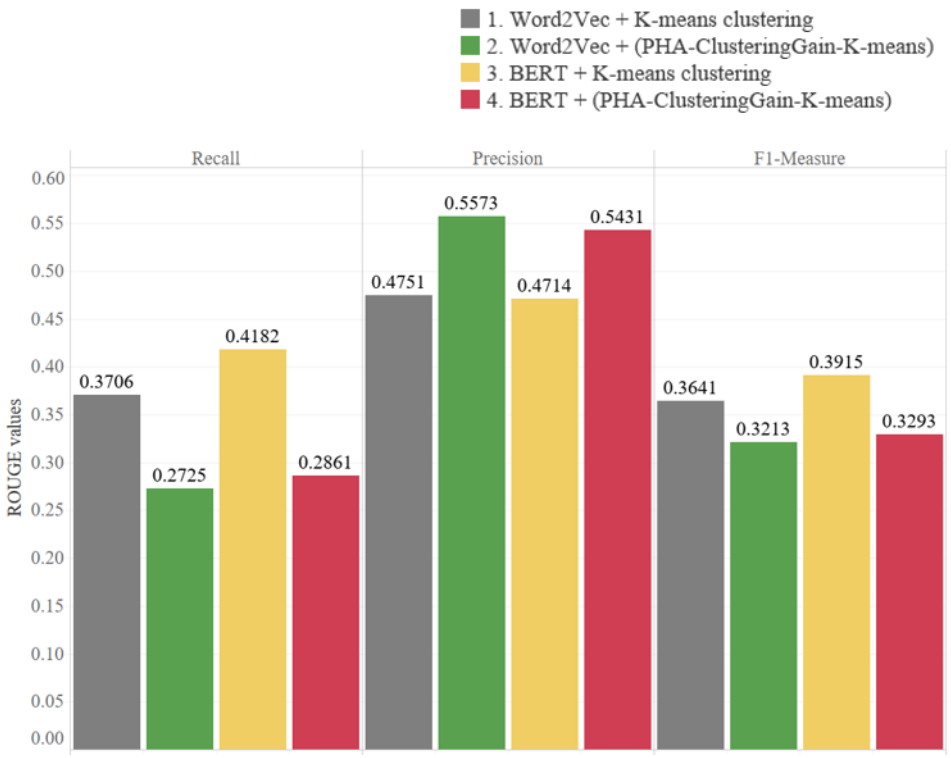


Figure 5.1: Bar-plot illustrating the performance measures of experiments

ROUGE which stands for Recall-Oriented Understudy for Gisting Evaluation is a method which helps with measures for spontaneously determining the quality of an automatically created summary by assessing it with other manually created summaries. The measures

consist counting of the number of units which are overlapping like n-gram based on n value, sequences of words and the pairing of word the among the automatically generated summary and the manual summary.

There are basically three measures which are generated from ROUGE evaluation metrics like Recall, Precision and f1-measure.The above Figure 5.1 illustrates the ROUGE scores which are obtained in the form of Recall, Precision and F1-measure from four different approaches of experiments. Beginning with the Recall measure, it can be observed from the figure that the recall value kept of changing drastically as different approaches of experiments were performed. Recall value helps in determining the amount or the percentage of sentences from reference manual summary that are identified by the algorithmic summary. It is calculated as the number of sentences found mutually among the algorithmic and reference summaries divided by the number of sentences in the reference summary. This basically gives importance to the reference summary while evaluation rather than the algorithmic summary which are generated using algorithms. As per the bar plot demonstrating the measures from various experiments, K-means clustering algorithm achieves much higher recall values when compare with the hybrid approach. Among those two experiments involving conventional K-means clustering, one using BERT model for feature embedding achieves maximum 41.82% of recall value.

Another important measure is the precision value, it helps in determining the percentage of correct sentences which are chosen by the reference manual summary and the automatically generated summary. This precision aspect focusses the number of n-grams in the algorithmic summary which are common with the reference summary. It is mathematically calculated as the number of sentences found in both algorithmic and reference summaries divided by the number of sentences in the algorithmic summary. This becomes highly critical to consider this measure when we try to produce summaries which are concise in nature. While comparing the most precise and outperforming algorithm from all these above four experiments, it is visible that there is not much drastic difference in the precision value among all the experiments. But it is clearly evident that the Hybrid approach of using "PHA-ClusteringGainK-means clustering" achieves higher precision value when compared with conventional approach. Among those hybrid approaches performed, the one using

word2vec as feature embedding method attains 55.73% as precision value. Hence this approach outperforms and provides more precise outcome than the one using BERT model for feature embedding.

Since the F1- measure is the computational combination of Recall and Precision, its importance differs as per the context of research. Based on the context of this study,it is best enough to compute the Precision and Recall values.

5.3 Hypothesis Evaluation

The objective of this research was to perform text summarization by implementing different unsupervised learning techniques like a hybrid clustering approach and a conventional clustering approach and also to determine the best approach which can deliver the better performance in terms of the precision. In order to accomplish the objective, the hypothesis of this research was structured as follows:

H0 : The application of PHA-ClusteringGain k-Means hybrid approach in text summarization will result in no precision increase over a conventional k-means clustering model.

The below Table 5.2 shows the precision scores which was obtained as part of evaluation using ROUGE-2 metrics from all the experimental approaches.

Experimental approaches	**Precision**
Word2Vec+K_means	0.4751
Word2Vec+PHA_Kmeans	0.5573
BERT+K_means	0.4714
BERT+PHA_Kmeans	0.5431

Table 5.2: Results for Hypothesis Evaluation

From the above table its well evident that irrespective of the approaches that are used for implementing the hybrid model of “PHA-ClusteringGain k-Means” and the conventional model of “k-means clustering”, both the approaches of using Word2Vec and BERT feature

embedding along with hybrid PHA-ClusteringGain k-Means algorithm achieved increase in the precision when compared with the conventional k-means clustering model. Among those hybrid approaches performed, the one using Word2Vec as feature embedding method attained 55.73% as maximum precision value.

Hence, the null hypothesis is rejected, and it can be concluded that the application of PHA-ClusteringGain k-Means hybrid approach in text summarization will result in precision increase over a conventional k-means clustering model.

5.4 How these results differ with previous researches

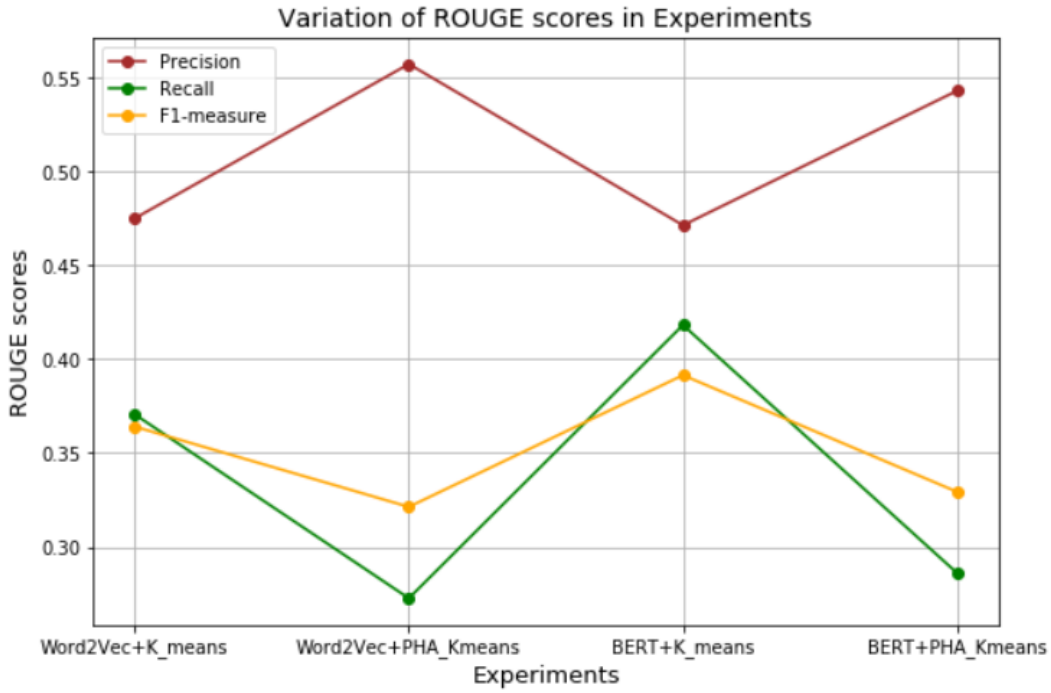


Figure 5.2: Line graph demonstrating the variation of ROUGE-2 scores in experiments

Most of the researches performed in the field of text summarization were using only conventional approaches of clustering and due to the availability of different types of approaches, these sorts of studies were implemented in numerous ways. But the approach used in this study is a lot more different and much more detailed into a comparative study between a hybrid clustering model and a conventional model. Basically, in this research study, four

different approaches of text summarization have been conducted by using various combinations of feature embedding technique like Word2vec /BERT model and hybrid/conventional clustering algorithms.

One of the similar research work from [Naveen, G., Nedungadi, P., (2014)], concluded that using PHAClusteringGain- KMeans hybrid clustering approach, by combining PHA method and kmeans method is more efficient and accurate than a conventional Hierarchical Agglomerative Clustering (HAC) algorithm. The results obtained from that study achieved a Precision score of 30.92% using ROUGE-2 method. The Figure 5.2 describes the variations of all the three ROUGE measures from different experimental approaches described in the above sections. It clearly shows the variations of inclined change in precision values when using hybrid approach of PHA-ClusteringGain-kmeans rather than conventional K-means algorithm. Irrespective of the type of feature embedding techniques, in both the experiments (Word2Vec+PHA_Kmeans,BERT+PHA_Kmeans), hybrid approach is outperforming than the conventional clustering approach and a maximum of 55.73% of Precision score is achieved in this research.

5.5 Strengths and Limitations of research

The main strength of this research study was its ability to precisely generate the precise summaries of the email content from the data. The hybrid clustering approach which was utilized has proved to outperform by achieving better class precision than the conventional clustering algorithm.

The approach for implementing hybrid model in this study could achieve a maximum precision score of 55.73%, which is more than 25% as that of an existing work implemented in a different approach.

At the beginning of this study, the dataset had over 5,00,000 emails records and a benchmark of 329 words was set as cutoff for training and testing data. This research also demonstrated that, not all the data is necessary to train the model to achieve the better performance. After doing the benchmark cutoff the data size could be scaled down from 5,17,401 to 79,565. Among these a lot more duplicate emails records were removed as part of data

cleaning before processing. This could help in reducing the time taken to train and test the data and thereby improving the processing speed.

One of the main limitation is rather challenging to get a public email data for researches and studies, due to privacy related issues and GDPR establishment. This is a drawback and a limitation especially for research since the studies cannot be conducted due to unavailability of public email datasets. An exemption to the beyond challenge is the Enron Corpus [Klimt, Yang, (2004)], in which this email data was made public after a legal investigation regarding the Enron Corporation.

Chapter 6

Conclusion

6.1 Introduction

This chapter concludes the dissertation as well as this research study and explains how the previously stated targets and objectives have been accomplished by the research study. The research objectives, designs and experimentations are repeated against the stated goals together with the findings. It is also extended to detail about the contributions and the impact made by this research. Finally, the areas of interest for future work along with some recommendations regarding this research are emphasized.

6.2 Research Overview

The research in this study is carried out to analyze the performance in terms of the precision of different Unsupervised learning approaches on the email data in implementing text summarization. The research started by reviewing the existing literature that was relevant for the topic, discussing various topics such as the traditional approaches in the field of text summarization, challenges in the field of text summarization, explaining various unsupervised learning algorithms and the feature embedding techniques that are well prominent in the existence, problems caused by the conventional k-means clustering and how the hybrid approach overcome these problems and improves the performance in the text summarization task. Following that, the research problem and the hypothesis was developed to direct the

study in achieving the research objectives. Hence, a quantitative research methodology was selected, and appropriate experiments were designed to guide the research. The proposed design was implemented using different approaches of Unsupervised learning algorithms along with various combinations of feature embedding techniques, then the results of all experimental approaches were obtained and evaluated using precision score from ROUGE-2 evaluation metrics. Eventually, the hypothesis evaluation is done using the achieved results from different experiments, in order to understand if the research objectives were attained.

6.3 Problem Definition

The study in this research was aimed to evaluate the performance in terms of precision from different approaches, using Unsupervised learning algorithms along with various combinations of feature embedding techniques in implementing text summarization task. In order to address this, the data from "Enron" email dataset was obtained and different approaches of Unsupervised learning algorithms combined with different feature embedding techniques were used to implement the text summarization task and to determine the finest approach that outperforms in terms of precision.

6.4 Design/Experimentation, Evaluation & Results

In this research, a quantitative and inductive research has been designed to address the research question. The CRISP-DM methodology is pursued throughout the study and Python programming using Jupyter Notebook/Google colab was used to implement the research.

The initial phase of this research was used to study the existing literature in the field of text summarization, then the corresponding research problem was analyzed and structured the research question. The necessary data for this study was obtained from "Enron" email data and the proper understanding about the data was completed. Following that, the phase of data preparation was performed, which comprised various data cleaning procedures like spell correction, NER for identification of name entities, abbreviation handling, case lowering, punctuation removal, tokenization and stemming/lemmatization. This pre-

processed data is then utilized to perform the text summarization task with various unsupervised learning techniques by incorporating different approaches of hybrid/conventional clustering algorithms along with various combinations of feature embedding techniques like Word2vec/BERT model. The results from all these experimental approaches was obtained and evaluated using ROUGE-2 evaluation metrics. This evaluation metrics provided with Recall, Precision and F1-measure values of all the experimental approaches. Since this research study is aimed to the evaluate all the experimental approaches and determine the best unsupervised learning approach by considering the performance in terms of precision, only the Precision score is being counted for performing the hypothesis evaluation and answering the research question. The final results revealed that irrespective of the feature embedding techniques and approaches being incorporated, the hybrid approach achieved increase in the precision when compared with the conventional k-means clustering model. Among those hybrid approaches performed, the one using Word2Vec as feature embedding method attained 55.73% as maximum precision value. Eventually, hypothesis evaluation is performed and concluded that the application of PHA-ClusteringGain k-Means hybrid approach in text summarization will result in precision increase over a conventional k-means clustering model and there by answering the research question.

6.5 Contributions and impact

Most of the text summarization literature made use of Unsupervised learning methods by using traditional word embedding techniques and conventional clustering algorithms. But only limited research studies were based on the Hybrid approach of Unsupervised learning methods such as PHA-ClusteringGain-Kmeans clustering algorithm, whereas this research makes use of this hybrid approach of PHA-ClusteringGain-Kmeans clustering algorithm to perform the text summarization task and hence contributing to the literature for the future research and studies.

By this research study, a text summarization approach with 55.73% of precision value when compared to the traditional and conventional models could be generated through this study. Hence, this approach can add on much more impact in the field of text summariza-

tion and Unsupervised learning to utilize the hybrid approaches rather than the traditional conventional approach.

6.6 Future Work & recommendations

This research has presented various approaches using Unsupervised learning methods to implement the task of text summarization. These approaches and models were built using the email data of "Enron" email dataset. As part of future work, a better and much more precise models could be generated using different kinds of data like Wikipedia data, documents, long paragraphs which are much cleaner. In this research the primary aim was to count only precision value based on the context of our research problem and this value was used for the answering the research question. As part of future work, either recall or f1 measure can also be focused and used to answer different contexts of research questions. As it is known that, this email dataset used in this research is a massive data with over 5,00,000 email records. If the same data is reused for any other research or study, it is recommended not to use the entire data. Instead, extract only those records which are necessarily enough as per the context of research. This will certainly help in reducing the processing time and speed up the work. In addition, based on the results from this research, it is highly recommended to use latest and finest feature embedding techniques so as to improve the precision of summaries and there by generating much efficient models.

6.7 Conclusion

This chapter concludes the research study and the experiments performed out to evaluate whether PHA-Clustering Gain k-Means hybrid approach out-perform a conventional k - means clustering model for email-based text summarization. This chapter included a brief overview of the research study that was carried out, sketched the experiments that were carried out and the attained results. Finally, presented some additional areas like contributions and impacts of this research along with some areas for future work as well as some recommendations to build similar experiments which are carried out as part of this research.

References

Agrawal, A., & Gupta, U. (2014). Extraction based approach for text summarization using k-means clustering. In *Proceedings of the international conference on information and knowledge management* (Vol. 4, p. 9–12).

Ando, R. K., & Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, *6*(2), 1817–1853.

Azevedo, A., & Santos, M. (2008). Kdd, semma and crisp-dm: a parallel overview. *Information Technology - IADS-DM*, *10*(1), 1-10.

Bengio, Y., Ducharme, R., & Vincent, P. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*(2), 137-1155.

Berger, A., & Mittal, V. (2000). Ocelot:a system for summarizing web pages. In *Proceedings of the 23rd annual international acm sigir conference, athens, greece* (p. 144-151).

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Cloze procedure: A new tool for measuring readability. *Journalism Bulletin*, *18*(4), 467–479.

Croft, W. B., Metzler, D., & Strohma, T. (2009). Search engines - information retrieval in practice. In (p. 75-125). Pearson Education.

Dai, A. M., & Le, Q. V. (2015). Semi-supervised sequence learning. In *Advances in neural information processing systems* (Vol. 3, p. 3079–3087).

Das, D. (2007). A survey on automatic text summarization single-document summarization. In *In proceedings of the international conference on information and knowledge management.* (Vol. 1, p. 1–31).

Dolan, W. B., & Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In *In proceedings of the third international workshop on paraphrasing* (Vol. 1, p. 89–105).

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on information and knowledge management* (p. 148-155).

Edmundson, H. (1969). New methods in automatic extracting. *Journal of the ACM*, *16*(2), 264–285.

Grossman, D. A., & Frieder, O. (2004, June). Information retrieval: Algorithms and heuristics. In *The kluwer international series on information retrieval* (p. 106-111).

Han, J., & Kamber, M. (2005). Data mining: Concepts and techniques. In (p. 100-135). Morgan Kaufmann.

Hinton, G., McClellanda, J., & Rumelhart, D. (1986). Distributed representations. In *In: Parallel distributed processing: Explorations in the microstructure of cognition.* (p. 113-135). MIT Press.

Hovy, E., & Lin, C. (2018). Automated text summarization in summarist. In *In i. mani m. maybury (eds.), advances in automatic text summarization* (p. 18-24).

Jafarpour, S., Burges, C. J., & Ritter, A. (2010). Filter, rank, and transfer the knowledge: Learning to chat. In *Advances in ranking* (p. 7-10).

Jain, A., & Dubes, R. C. (1998). Algorithms for clustering data. In (p. 4-125). Prentice Hall.

Jain, A., & Murty, M. (1999). Data clustering: a review. In *Acm comput. surv.)* (p. 264-323).

Jardine, N., & Rijsbergen, C. J. (1971). The use of hierarchical clustering in information retrieval. In *Information storage and retrieval* (Vol. 7, p. 217–240).

Jernite, Y., Bowman, S. R., , & Sontag, D. (2017). Discourse-based objectives for fast unsupervised sentence representation learning. *CoRR*, *2*(2), 758-786.

Jones, K., & Willett, P. (1997). Readings in information retrieval. In (p. 25-90). Morgan Kaufmann Publishers Inc.

Jung, Y., & Park, H. (2002). A decision criteria for the optimal number of clusters in hierarchical clustering. In *Kluwer academic publishers)* (p. 1-3).

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In *In advances in neural information processing systems* (Vol. 11, p. 3294–3302).

Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. In *Proceedings of the 18th annual international acm sigir conference, seattle, wa* (p. 68-73).

Kushmerick, N., & Lau, T. (2005). Automated email activity management: An unsupervised learning approach. In *Proceedings of 10th international conference on intelligent user interfaces, acm press* (p. 67-74).

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *In international conference on machine learning* (Vol. 11, p. 1188–1196).

Lee, J., Park, S., Ahn, C., & Kim, D. (2009, Jan). Automatic generic document summarization based on non-negative matrix factorization. In *Information processing and management* (Vol. 45, p. 10-24).

Lee, J., & Park.S. (2009,). Automatic generic document summarization based on non-negative matrix factorization. In *Information processing and management, sciencedirect)* (p. 20-34).

Lewis, D. D., & Knowles, K. A. (1997). Threading electronic mail: A preliminary study. In *Ieee international conference on control system, computing and engineering)* (Vol. 33, p. 209-217).

Lin, C. (2004). Rouge: A package for automatic evaluation of summaries. In *Proceedings of the international conference on information and knowledge management)* (p. 1-3).

Lu, Y., & Wan, Y. (2013). Pha: A fast potential-based hierarchical agglomerative clustering method. In *Pattern recognition, sciencedirect, volume 46)* (p. 1227-1239).

Lu, Z., & Li, H. (2013). A deep architecture for matching short texts. In *Advances in neural information processing systems* (p. 1367–1375).

Luhn, H. (1958). The automatic creation of literature abstracts. In *International conference on computer and information science* (p. 159-165).

McKeown, K., Shrestha, L., & Rambow., O. (2007). Using question-answer pairs in extractive summarization of email conversations. In *In proceedings of the international conference on computational linguistics and intelligent text processing* (Vol. 3, p. 542–550).

Mihalcea, R. (2004). Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *In proceedings of the acl 2004 on interactive poster and demonstration sessions (acldemo '04),association for computational linguistics)* (p. 1-3).

Mikolo, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *In proceedings of workshop at iclr* (Vol. 1, p. 89–152).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *In advances in neural information processing systems* (Vol. 26, p. 3111–3119).

Mnih, A., & Hinton, G. E. (2009). A scalable hierarchical distributed language model. In *In advances in neural information processing systems* (Vol. 21, p. 1081–1088).

Montanes, E., Diaz, I., Ranilla, J., Combarro, E. F., & Fernandez, J. (2005). Scoring and selecting terms for text categorization. In *2005 3rd ieee international conference on computer and communications (iccc)* (p. 40-47).

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, *26*(4), 354–359.

Naveen, G., & Nedungadi, P. (2004). Query-based multi-document summarization by clustering of documents. In *Proceedings of the 2014 international conference on interdisciplinary advances in applied computing* (Vol. 4, p. 3-9).

Nenkova, A., & Bagga., A. (2003). Facilitating email thread access by extractive summary generation. In *In proceedings of the recent advances in natural language processing conference* (Vol. 5, p. 85–106).

Newman, P., & Blitzer, J. (2003). Summarizing archived discussions: a beginning. In *In proceedings of the international conference on intelligent user interfaces* (Vol. 2, p. 273–276).

Ng, R., & Han., J. (1994). Efficient and effective clustering methods for spatial data mining. *VLDB Conference*, *6*(1), 226-258.

Padmakumar, A., & Saran, A. (2016). Unsupervised text summarization using sentence embeddings. In *Proceedings of the international conference on information and knowledge management* (p. 9–12).

Pan, S., & Yang, Q. (2010, Oct). A survey on transfer learning. In *Ieee transactions on knowledge and data engineering* (Vol. 22, p. 1345–1359).

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors forword representation. In *In empirical methods in natural language processing (emnlp)* (Vol. 26, p. 1532– 1543).

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations* (Unpublished doctoral dissertation). In NAACL.

Radev, D. R., Hovy, E., & McKeown, K. (2002). Introduction to the special issue on summarization. In *Proceedings of the international conference on computational linguistics* (Vol. 28, p. 399–408).

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding with unsupervised learning. *Technical report, OpenAI*, *1*(1), 41-62.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. In *In proceedings of the 2016 conference on empirical methods in natural language processing* (Vol. 3, p. 2383–2392).

Rambow, O., Shrestha, L., Chen, J., & Lauridsen, C. (2004). Summarizing email threads. In *In human language technology conference of the north american chapter of the association for computational linguistics* (Vol. 2, p. 21–61).

Salton, G. (1983). An introduction to modern information retrieval. In (p. 15-85). McGraw Hill.

Sang, E. F. T. K., & Meulder, F. D. (2003a). *Introduction to the conll-2003 shared task: Language-independent named entity recognition.* (Unpublished doctoral dissertation). NAACL.

Sang, E. F. T. K., & Meulder, F. D. (2003b). Language-independent named entity recognition. *CoNLL*, *2*(4), 72-86.

Scheffer, T. (2004). Email answering assistance by semi-supervised text classification. In *Proceedings of sixth international conference on document analysis and recognition* (Vol. 8, p. 481-493).

Schutze, H., & Silverstein., C. (1997). Projections for efficient document clustering. *ACM SIGIR Conference*, *1*(2), 75-88.

Segal, R. B., & Kephart, J. O. (1999). Mailcat: an intelligent assistant for organizing e-mail. In *Pro-ceedings of the third international conference on autonomous agents* (p. 276-282).

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *In proceedings of the 2013 conference on empirical methods in natural language processing* (Vol. 8, p. 1631–1642).

Taylor., W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Bulletin*, *30*(4), 415–433.

Tombros, A., Villa, R., & Rijsbergen, V. (2002, July). The effectiveness of query-specific hierarchic clustering in information retrieval. In *Information processing and management, sciencedirect,* (Vol. 38, p. 559-582).

Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. In *In proceedings of the 48th annual meeting of the association for computational linguistics* (Vol. 10, p. 384–394).

Turney, P. (2000). Learning algorithms for key phrase extraction. In *Proceedings of the special interest group on information retrieval* (Vol. 2, p. 303-336).

Ulrich, J., Murray, G., & Carenini, G. (2008). A publicly available annotated corpus for supervised email summarization. In *In proceedings of the aaai email workshop* (p. 77-87).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin., I. (2017). Attention is all you need. In *In advances in neural information processing systems* (Vol. 1, p. 6000–6010).

Vel, O. d., Anderson, A., Corney, M., & Mohay, G. (2001). Mining e-mail content for author identification forensics. In *In proceedings of the annual international acm sigir conference* (Vol. 30, p. 55-64).

Velmurugan, T. (2004). Performance based analysis between k-means and fuzzy c-means clustering algorithms for connection oriented telecommunication data,. In *Applied soft computing, sciencedirect, volume 19* (p. 134-146).

Voorhees, E. M. (2002). The philosophy of information retrieval evaluation. In *Proceedings of the special interest group on information retrieval* (p. 355–370).

Wan, X. (2007). Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In *Proceedings of the 2013 conference on empirical methods in natural language processing. association for computational linguistics* (p. 552–559).

Wang, H., Lu, Z., Li, H., & Chen, E. (2013). A dataset for research on short-text conversations. In *Proceedings of the 2013 conference on empirical methods in natural language processing. association for computational linguistics, seattle, washington, usa* (Vol. 1, p. 935–945).

Wirth, R., & Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. *Information Technology*, 11.

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the special interest group on information retrieval* (p. 42-49).

Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international acm sigir conference on research in information retrieval.* (p. 334–342).

Appendix A

Appendix 1

1. Format of data at the initial stage after separating the body content for processing from the message content of the Enron email dataset:

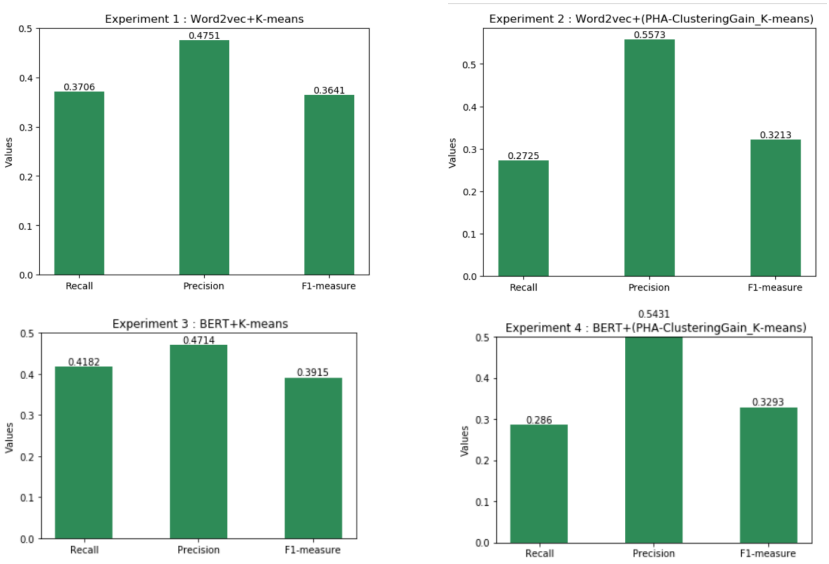
file	message	body
allen-p/_sent_mail/1.	Message-ID: <18782981.1075855378110.JavaMail.e...	Here is our forecast
allen-p/_sent_mail/10.	Message-ID: <15464986.1075855378456.JavaMail.e...	Traveling to have a business meeting takes the...
allen-p/_sent_mail/100.	Message-ID: <24216240.1075855687451.JavaMail.e...	test successful. way to go!!!
allen-p/_sent_mail/1000.	Message-ID: <13505866.1075863688222.JavaMail.e...	Randy, Can you send me a schedule of the sal...
allen-p/_sent_mail/1001.	Message-ID: <30922949.1075863688243.JavaMail.e...	Let's shoot for Tuesday at 11:45.

2. Format of data after separating emails messages which are having word count greater than 329 words (word count mean value):

file	message	body	forward_flag	token_count
allen-p/_sent_mail/116.	Message-ID: <25140503.1075855687800.JavaMail.e...	Liane, As we discussed yesterday, I am conce...	0	421
allen-p/_sent_mail/130.	Message-ID: <31434120.1075855688116.JavaMail.e...	George, Below is a list of questions that Ke...	0	458
allen-p/_sent_mail/2.	Message-ID: <5468446.1075855378133.JavaMail.ev...	Outlook Migration Team@ENRON 05/11/2001 01:49 ...	0	664
allen-p/_sent_mail/347.	Message-ID: <30467968.1075855723641.JavaMail.e...	Outlook Migration Team@ENRON 05/11/2001 01:49 ...	0	686
allen-p/_sent_mail/421.	Message-ID: <24048786.1075855725309.JavaMail.e...	Reagan, Thank you for the quick response on t...	0	393

3. Format of processing data during the data cleaning process:

file	message	body	forward_flag	token_count	spell_corrected	spell_abb_corrected	final_corrected
allen-p/_sent_mail/116.	Message-ID: <25140503.1075855687800.JavaMail.e...	Liane, As we discussed yesterday, I am conce...	0	421	Liane, As we discussed yesterday, I am conce...	Liane, As we discussed yesterday, I am conce...	Liane, As we discussed yesterday, I am conce...
allen-p/_sent_mail/130.	Message-ID: <31434120.1075855688116.JavaMail.e...	George, Below is a list of questions that Ke...	0	458	George, Below is a list of questions that Ke...	George, Below is a list of questions that Ke...	George, Below is a list of questions that Ke...



4. Format of test data which are used for the feature embedding and modelling process:

email	number of clusters	manual summery
"Liane, As we discussed yesterday, I am conc...	5	At the time of these trades, offers for physic...
George, Below is a list of questions that Ke...	8	George, Below is a list of questions that Keit...
"Reagan, Thank you for the quick response on ...	9	What type of floor joist would be used?\r\n\r\...
"George, I am back in the office and ready to...	7	Specifically that the costs of our project are...
Lucy, Please fix #41 balance by deleting the ...	5	The other questions I had about last week\s re...
"George, The probability of building a house ...	6	George, The probability of building a house th...

5. Format of results obtained as part of evaluation:

email	number of clusters	manual summery	summery	recall	precision	f1_measure
"Liane, As we discussed yesterday, I am conc...	5	At the time of these trades, offers for physic...	For San Juan Gas With The Intent To Distort T...	0.327586	0.557185	0.412595
George, Below is a list of questions that Ke...	8	George, Below is a list of questions that Keit...	George, Below Is A List Of Questions That Ke...	0.787611	0.154313	0.258065
"Reagan, Thank you for the quick response on ...	9	What type of floor joist would be used?\r\n\r\...	Below Is A List Of Questions On The Specs ...	0.546980	0.256289	0.349036
"George, I am back in the office and ready to...	7	Specifically that the costs of our project are...	Million. Million Land ...	0.297045	0.459135	0.360718
Lucy, Please fix #41 balance by deleting the ...	5	The other questions I had about last week\s re...	Why I Spoke To Jeff Smith. He Was Not Tryi...	0.217647	0.412639	0.284981
"George, The probability of building a house ...	6	George, The probability of building a house th...	This Plan Still Has The Same Number Of Rooms...	0.332253	0.508685	0.401961