Republic of Iraq
Ministry of Higher Education and
Scientific Research
University of Babylon
College of Information Technology

Extractive Topic Summarization Based on Improved Lexical Chains Sentences and Neural Network

A Dissertation

Submitted to the Council of the College of Information Technology, University of Babylon in Partial Fulfillment of the Requirements for the Doctor of Philosophy Degree in Information Technology / Software

By

Marwan Badran Mohammed Rasheed

Supervised by

Asst. Prof. Dr. Wafaa Mohammed Saeed Hamzah

2022 D.C. 1442 A.H.

بسم الله الرحمن الرحيم

وَلَسَوْفَ يُعْطِيكَ رَبُّكَ فَتَرْضَىٰ

حامد ١٣٦٦

صدق الله العظيم

سورة الضحى الاية (5)

Dedication

To

My Beloved Parents,

My Wife,

My Brothers,

All Friends,

And everyone who supported me with

Love and respect

Marwan Badran Mohammed

Acknowledgments

Thanks, and gratitude to Allah who gave the ability to accomplish such incomplete work and all that was done with the support and success of God.

Gratitude to all of my family, especially my mother, and my wife, were the most important motivation for them patience and encouragement, which enabled me to complete my research. Also, I express my thanks to my teachers, Prof.Dr. Ban Nadeem, and Prof.Dr. Balal Asmaeal to provide Tips that serve in the completion and success of this work.

I want to express my sincere gratitude and appreciation to my supervisor (Assent Prof., Dr. Wafaa Mohammed Saeed) for her invaluable guidance, supervision, and tireless efforts during my studies.

In addition, special thanks to the Head of the Software Department, and all College of Information Technology faculty members for their sincere efforts and support in all directions. For their help in letting me make use of most of the theoretical and practical enhancement.

Marwan Badran Mohammed

Abstract

Increasing growth in the volume of digital data of documents performed the difficulty of accessing important information. The solution is using automatic summarization systems that aim to extract important information in a short time. The work of these systems is usually to extract a single summary from a single document or multi-documents.

This dissertation presents several directions. The first is to propose a new algorithm that coined the name Develop Clustering Algorithm (DCA) to collect unlabeled data and put it in appropriate groups. The second is the creation of a lexical chain based on similar semantic sentences or a similar number of words among sentences coined the name Lexical Chain Sentences (LCS), which is different from the traditional Lexical Chain Word (LCW) that works based on words. The third is to propose a set of features to extract important sentences and easy to understand. The fourth is building a Backpropagation Multi-layer Perceptron Neural Network (BMPNN) to find a sentence score. The fifth is using the Random oversampling (ROS) method and its effective role in rebalancing the data during the training process in BMPNN.

Finally, the problem of reordering sentences in the candidate summary according to the importance of the sentence is solved by depending on the date and in addition to three conditions taken into consideration to ensure the accuracy rearrangement process. This work used two datasets of interest in articles of news. Dataset one is the Document Understanding Conference (DUC 2002), and Dataset two has been created manually from the news documents collected by us for experiments.

The results showed that the performance of the proposed DCA algorithm has generally outperformed the hierarchical clustering algorithm by the number of clusters generated, and also, on the K-mean algorithm by evaluation results produced from the Davies Bouldin Index (DBI) metric. Regarding the evaluation of the candidate summary. This dissertation used three measures of the Recall-Oriented Understudy for Gisting Evaluation (Rouge) family measures, Rouge-1, Rouge-2, and Rouge-L to evaluate the candidate summary. The results of the evaluation of Rouge-1, Rouge-2, and Rouge-L measures showed that the candidate's summary is very close to the golden summary in terms of matching sentences, and it achieved promising results. The average accuracy for rouge metrics above in all topics in the DUC 2002 dataset is (0.81, 0.75, and 0.81) respectively when the reference word summary is 200 words; (0.76, 0.69, and 0.76) respectively when the reference word summary is 400 words; and (0.78, 0.72, and 0.78) respectively when the average reference word summary is between 200 and 400 words. While the results F-score for these measures above in second dataset for three topics are {(0.76,0.64, and 0.76), (0.69,0.62, and 0.69), and (0.96,0.92,0.96)} respectively when the reference word summary is (390,114, and 518) words respectively

Declaration Associated with this Dissertation

Some of the works presented in this dissertation have been published as listed below.

1) New Algorithm for Clustering Unlabeled Big Data. Indonesian Journal of Electrical Engineering and Computer Science (IJEECS).

   (Scopus – Q3)- Published

2) Cohesive Summary Extraction from Multi-Document Based on Artificial Neural Network. 7th International Conference On Contemporary Information Technology And Mathematics (ICCITM), and Published In The ICCITM Conference Proceedings In IEEE Xplore Digital Library.

   (Scopus) – Published

Table of Contents

List of Tables

List of Figures

List of Algorithms

List of Abbreviations

Abbreviation	Meaning
AF	Activation Function
ANN	Artificial Neural Network
BIRCH	Balanced Iterative Reduction And Clustering Utilizing Hierarchies
BMPNN	Backpropagation Multi-Layer Perceptron Neural Network
CS	Cosine Similarity
CSO	Cat Swarm Optimization
CURE	Clustering Using Representatives
DAN	Deep Averaging Network
DBCL	Density-Based Clustering
DBI	Davies Bouldin Index
DCA	Develop Clustering Algorithm
DENCLUE	Density-Based Clustering
DL	Deep Learning
DUC	Document Understanding Conference
ELU	Exponential Linear Unit
ESV	Embedding Sentences Vector
FFNN	Feed-Forward Neural Network
GA	Genetic Algorithm
HC	Hierarchal Clustering
HF	Headline Feature
HSA	Harmony Search Algorithm
KDD	Knowledge Discovery In Databases
LCLT	Lower Case Letters Technique
LCS	Lexical Chain Sentences

LCW	Lexical Chain Word
MLP	Multi-Layer Perceptron
MLPBNN	Multilayer Perceptron Backpropagation Neural Network
MLPNN	Multilayer Perceptron Neural Networks
MOO	Multi-Objective Optimization
NBOW	Neural Bag-of-Words
NC	Number Cluster
NF	Negative Feature
NLP	Natural Language Processing
NNWF	Number Noun Word Feature
OPTICS	Ordering Points To Determine The Clustering Structure
PF	Positive Feature
POS	Part Of Speech
PRelu	Parametric Rectified Linear Unit
PRT	Punctuation Removal Technique
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Networks
ROS	Random Over Sampling
ROUGE	Recall-Oriented Understudy for Gisting Evaluation Metric
SAF	Sigmoid Activation Function
SELU	Scaled Exponential Linear Unit
SLF	Sentence Length Feature
SNF	Sentence Numerical Feature
SPF	Sentence Position Feature
SSD	Sentence Sense Disambiguation
STT	Sentence Tokenization Technique
SVR	Support Vector Regression
TF Hub	Tensorflow Hub

TLBO	Teaching–Learning Based Optimization
TM	Text Mining
TP	True Positive
USE	Universal Sentence Encoder
w2v	Word Two Vector
WEAT	Word Embedding Association Tests
WSD	Word Sense Disambiguation
WSRT	Whitespace Removal Technique
WTT	Word Tokenization Technique

CHAPTER ONE

OVERVIEW OF THE TEXT SUMMARIZATION

1.1 Introduction

Internet's exponential expansion and the availability of a massive amount of online information makes identifying relevant content that satisfies user demands has become extremely difficult. This has sparked a race to develop technologies for automatic document summarizing. This race is crucial not only for professionals who need to obtain information quickly, but also for huge search engines such as Google, Yahoo, AltaVista, and others [1]. To address this issue, automatic text summarizing systems are required, which construct a summary from the specific text document(s) [2]. Consequently, automatic multi-document summarizing can be defined as the process of constructing a shorter version of a given text from several but related sources in such a way that the key information conveyed by the documents is retained [3].

Usually, many text mining applications require text document summarization in order to provide a succinct overview of a huge document or group of papers on a topic [4]. The concept of text summarization is the acquisition of a subset of influential information to portray the entire topic as briefly as possible. Text summarization is defined as the task of compressing some text into a shorter version of the original that contains most relevant and vital information, which may result in informational loss [5].

Text summarization can be used to create summaries of medical records, meteorological data, and the latest news, summarize product reviews, student responses to post-class questionnaires, collections of news articles on a specific topic, and among other things [6].

The Summary classification is based on whether the result is Generic or Query-based. The generic text summary gives readers a general sense of what the document is about without requiring any prior knowledge, whereas the query-based summary is the information offered in a query-relevant summary that should be relevant to a certain query or topic [7]. A query-based summary displays the data related to the original search query. A generic summary, on the other hand, gives an overview of the article's content, the text summarization is split into three following approaches [8] [9]:

1. **The Extractive summarizing approach** is the process of extracting relevant data from documents and literally summarizing it [5]. Figure 1.1 depicts the architecture of the extractive text summarization system, which is made up of three major components. The input text is pre-processed in the first half. The second step involves post-processing (for instance rearranging sentences that have been extracted, Pronouns are replaced with their antecedents, relative temporals are replaced with actual dates, and so on). The processing tasks are described in three steps [10]:
    a. Constructing an appropriate to facilitate text analysis, a representation of the input text is used (e.g. N-gram, bag-of-words, graphs, etc.).
    b. Sentence scoring: evaluating sentences based on their representation in the input text.
    c. Extraction of high-scoring sentences: The summary is constructed by concatenating the most important sentences from the input documents. The length of the created summary is determined by the preferred compression (or reduction) rate, which is limited by a length limit or threshold that keeps the generated sentences in the same sequence as the original text.

Typically, the process of extractive summarization in current computer systems consists of three stages. The creation of an intermediate representation of the original text is part of the initial phase of the process. The second stage is to assess the summary and construct a scale for evaluating the sentences. As a result, the key sentences can be identified in this step. The third phase involves the selection of coherent sentences for the building of the summary. This stage follows a difficult style and often involves the selection of the sentences that have the highest score by evaluation. [11] [12].

The extractive method is quicker and easier to use than the abstraction method. Due to the extraction of sentences directly, this method produces more accuracy since the summary is read using the same vocabulary as the original [13]. Furthermore, extracting summarization is considerably more practical for modern summarizing algorithms, which focus on finding the most important elements of a piece of text to provide a short summary. However, the produced text is frequently incoherent. The reader can, however, develop an opinion on the original content [14]. In addition, one of the most difficult aspects of "extractive summarizing" is identifying key sentences from the original source and using them verbatim in the summary. It entails combining source sentences to form a summary [5]. Another issue is that this method differs significantly from how human experts compose summaries [15].

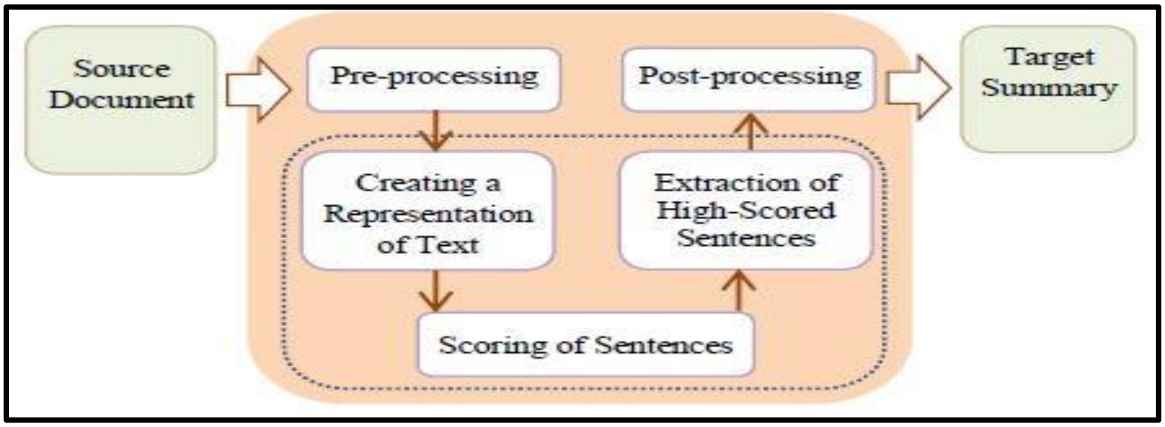

Figure 1. 1 An extractive text summarization system's architecture in general. [10]

2. **The abstractive summarization approach** necessitates a more in-depth examination of the input text. Abstractive summarization uses different words that do not exist in the original document to construct the summary, expressing the ideas of that document [5] [16]. thereby making it more complex [5] ". This approach is more efficient than extractive summarization because it produces summaries that are more human-like [16]. Also "Abstract summarization" algorithms typically generate original sentences depending on the document's content. However, these strategies are significantly harder to execute than extractive summarization techniques in general. Therefore, numerous approaches attempt to combine the principles of extraction and abstraction in their summaries [17].
3. **The hybrid abstractive and extractive techniques**. It is usually divided into many stages [18]: Preprocessing, extractive phase: extract the most significant sentences from the text, then apply abstractive procedures and techniques in the abstractive phase, and post-processing: check for errors. The two approaches, extractive and abstractive, become complementary as a result of this strategy, and the performance of summarization as a whole enhances. However, because the created summary is based on the extracts rather than the actual text, it produces a lesser quality abstractive summary than the pure abstractive technique. Because the abstractive approach is highly difficult and requires substantial Natural Language Processing (NLP), the research community is focused more on the extractive approach than the abstractive approach, employing various methodologies and strategies to generate more coherent and relevant summaries [10].

The Summarization Algorithm can be either supervised or unsupervised. A training step is required for the supervised algorithm, which requires annotated training data. Unsupervised algorithms do not need the training data [19].

The most significant advantage of text summarization is that it can reduce the amount of time a user spends reading. While maintaining a low level of repetition, a fantastic text summary system should reproduce the diverse theme of the content. Several various aspects of summarizing must be taken into account [20] :

- **Construct**: A natural language is a semantic representation that shows the schema text and important points of a textual document when one wants to make an abstract summarization. In contrast, an extracted summary includes parts of the original text like keywords, phrases, paragraphs, full sentences, and/or concise sentences.
- **Type:** A summary is a short story that explains the main points of a document in a way that is easy to read. This includes a set of keywords, a title, an abstract, an extract, a goal-focused abstract, an index, or a Table of contents.
- **Purpose:** The purpose is to provide an overview summary of the entire document. While a query-relevant summary provides content that corresponds to the user query or user model, a goal-focused summary provides information that corresponds to a specific purpose.
- **The number of summarized documents:** Single summary can be made for just one document, or a summary can be made for multiple documents to give an overview of all of them at the same time
- **Document length:** knowing how much repetition is contained in a single document is possibly made by document length, so the newswire articles are recaps of an affair. The least amount of frequencies are used in these publications. Official documents, on the other hand, are written to demonstrate a certain intention, which is then re-explained in the outline.
- **User goal:** When a user wants to find specific information, the objective is a robust informative summary construction that takes into account the source documents.

- **Genre :** the texts contain linguistic and structural features that are useful for establishing a summary. There are numerous types of documents based on their genres, such as news articles, opinion pieces, letters and memoranda, emails, scientific documents, novels, web pages, legal decisions, and speech transcripts.
- **Presentation:** The summary can be displayed in either text-only or in text-with-hyperlink to the source text(s). Both of these formats are considered to be text-based formats. Generally speaking, any summary may include one or more of the following elements: keywords, phrases, sentences, and paragraphs, to name a few. The length of the summary could be made longer or shorter by giving more context about each sentence summary or else part in the state of the text extractive summaries.
- **Source language:** When retrieving multilingual information by parts, the language of the source document set may differ from the language of the output document set [20]. When it comes to language-based classification, A summarization system is considered to be monolingual when both the origin and goal documents are written in the same language. On the other hand, the original content is written in a variety of languages (for example, English, Arabic, and French) and this system is claimed to be multilingual because the summary is generated in various languages as well. The cross-lingual is intended for the original text (for instance, English) and the summary (e.g. Arabic or French) [10].
- **Summary Content Classification:** Indicative or Informative. An indicative summary simply conveys the main thought of the original text. So it determines the input text's content (i.e. what topics are addressed) and alerts the user. An indicative summary informs users about the input text's scope, allowing them to select whether or not to read it. A comprehensive summary, on the other hand, includes all relevant material and concepts from the original text. An informative summary covers the major points of the original material without elaborating [10].

This dissertation aims to improve and develop the strategy of the topic summary system from multiple documents to extract the important sentences and form a coherent and understandable summary that covers the topic and matches the opinion of the reader or writer. Thus, it introduces the proposed (enhances) topic summarization system to work on an extractive approach. Characterize the proposed system that it works as semi unsupervised which means the collecting between clustering and classification algorithm. This proposed system includes several modern directions. the first is creating a lexical chain based on sentences (LCS). The second is to propose an MB clustering algorithm that contributes to establishing LCS. The third is proposing a new combination of feature extraction. The fourth is building a backpropagation neural network for finding sentences scores. The last one is choosing important sentences from multiple documents that correspond with the human expert's opinion and reordering the candidate sentences summary to produce a cohesive candidate summary by preparing some criteria.

1.2 Motivation

Text Summarization has become an important study issue due to the rising availability of online textual databases and the advent of the Internet. Furthermore, the writer's knowledge of the subject and the effect of subjective opinion reflect in the summary, manually writing summaries is a time-consuming and error-prone operation Thus, several important issues are discussed in summarizing the text in this treatise: How can we create a lexical dealing with an entire sentence rather than words to overcome the ambiguity of words?. How can we produce cohesive, consistent, and understandable summary events that cover the topic of the article while consuming less time?.

1.3 The Statement of the Problems

- Extract best cohesive candidate summary for topic by multi-documents.

- Find of important sentences and extracted of the text.
- Grouping the important sentences into candidate summary which coverage whole topic with avoid redundant of sentences .
- Reordering important sentences in candidate summary Through chronological order to be create a summary concept and cohesive.

During the implementation of the suggested system or the current dissertation, various obstacles have arisen. They will be mentioned in the following:

1. Extracting a single summary from a group of documents (for example, multiple documents). It is a difficult task because the input is various documents indicating the same topic.
2. The dataset employed in this suggested system is unbalanced, with the number of significant sentences being lower than the number of nonimportant sentences. As a result, one of the challenges in classification problems is to train a classifier that reliably predicts which class an observation belongs to. Using the Random oversampling (ROS) strategy to overcome the unbalanced dataset problem to assure the classifier's accuracy and avoid biasing it to one category exclusively. This strategy proved successful in resolving the problem.
3. Choosing a set of features extraction with the intention of helping to access important sentences by analyzing the results effects of these features on sentences is a challenge.
4. Calculating a sentence's score based on the appropriate algorithm, thus how choosing an active algorithm for this task achieves high accuracy is a challenge.
5. When applying reordering candidate sentences to achieve cohesive, consistent, and understandable summary events that cover the topic observed that not all documents in the dataset contain a date. In addition, For the same issue, not all documents include a time, such as (am or pm). also, a title is missing from several

of the documents. Finally, several of the documents on the same subject have the same date and time. Thus, for success in applying the reordering procedure for candidate sentences summary, added the token "$TEMP$" for document\s that do not contain a date. Also, added the token "Dd" to document\s that lack a timestamp ($i.e\ am\ or\ pm$),and added token "2400" represent clock missing. Documents that lack a title for the same topic take the title from the previous document and add it to these documents, and then use one of the extraction features proposed for rearrangement of the sentences in documents that have a date and time similarity to the same topic.

1.4 Aim and Objective of Dissertation

The aim of dissertation produces cohesive summary from multi-documents by focus on two key criteria. The first is coverage such that the candidate sentences for summary must cover whole topic. the second is variety such that this candidate summary must contain on variety sentences carry different events coordinated to the same topic. Thus, the suggested system selects different sentences that are important to cover the core topic, resulting in a high-quality summary. The objectives of this research can be described as follows:

1. This work used two datasets. The first DUC 2002 dataset. Another dataset collected from internet.
2. Create a new algorithm coined Developed clustering algorithm to work on the clustering concept to deal with numeric data which is unlabeled. This algorithm does not require the initial number of cluster prior because it determines the number of clusters automatically, and help the user collect similar objects by putting them in the cluster.

3. Evaluating the outputs of this proposed algorithm is by using one of the internal clustering measures, which is the Davies Bouldin Index (DBI) metric, and comparing the performance efficiency between this algorithm and two of the algorithms clustering.
4. Building lexical chain based on sentences coined Lexical Chain Sentences (LCS) interested to collect whole sentences rather than lexical chain based on words to overcome word ambiguity problems by depending on close sentences in the semantic or the number of words while being careful in not redundancy the sentence in more of a chain.
5. Creating a new mix of feature extraction techniques collected from some earlier studies to analyze sentences by extracting features to explore important sentences.
6. Building a multilayer perceptron backpropagation neural network (MLPBNN) to compute a score for each sentence.
7. Making a cohesiveness summary by retaining the major sentences that indicate the core topic of the document collection and removing the irrelevant, superfluous sentences from the entire collection.
8. Use Rouge metrics to evaluate candidate summary with the golden summary existence in the dataset used.

1.5 The Contributions of the Dissertation

The aim of dissertation will assist the reader in reading fewer data in less time. Because our proposed system can read a group of articles on a certain topic at the same time, deconstructing them and identifying the main concepts from the raw text in less time, it will write a decent, unified cohesive summary before the person can look at the articles. The following is a list of the dissertation's contributions:

1. Improved clustering of lexical chain sentences to satisfy the cohesion property of the text to distinguish between important sentences and those that are not. In this case,

the text will be separated into clusters/chains. Each chain consists of a group of sentences related to the same topic.

2. Using a backpropagation multi-layer perceptron neural network to find a score for each sentence.
3. Suggesting a new type of cohesion in order to determine a tightly connected text by creating lexical cohesion based on relating sentences instead of using traditional lexical based on relating words to produce a cohesive summary for multi-document as a new model. By discovering the effect of the context in obtaining the appropriate meaning (sense) of the word found within a specific region (the sentence) to obtain the exact meaning and to reduce the comparisons with all the words. Consequently, generate lexical chains which provide the cohesion text feature.

1.6 Related Works

There various extraction-based techniques have been proposed for text summarization.

1. In 2015 authors, Saleh and Kadhim in [1] focused on three issues which are content coverage, frequency, information, and length. These issues together make the global summary problem one of the most difficult tasks. They proposed a multi-text summary model for extracting text based on the Genetic Algorithm (GA) where they designed the problem as a separate optimization problem and the specific fitness function was designed for the purpose of effectively dealing with the proposed model. After that, a binary-encoded representation together with a heuristic mutation and local repair operators were proposed to characterize the adopted GA. Applied experimentally on ten topics from Document Understanding Conference (DUC) 2002 datasets (d061j through d070f). The proposed method was measured using the Recall-Oriented Understudy for Gisting Evaluation (denoted by the ROUGE evaluation metrics).

2. In 2016 the authors Saleh and Kadhim [21] proposed two models that were then coupled and defined as a Multi-Objective Optimization (MOO) problem. Heuristic perturbation and heuristic local repair operators were proposed and injected into the adopted evolutionary algorithm. Because their proposed work regards balancing the two significant objectives: the first, content coverage, and the second, diversity when generating summaries from a collection of text documents.Assessment of the proposed models was performed using document sets supplied by Document Understanding Conference (DUC) 2002 and a comparison is made with other state-of-the-art methods.
3. In 2018, Afsharizadeh, and et al [22] introduced a query-oriented text summarization technique that is proposed by extracting the most informative sentences, such that Ahuja's proposed method [23] has been improved by the addition of several appropriate query-based features. In this search, 11 of the finest attributes from each text are extracted. The feature set is enhanced with a variety of other appropriate features in order to identify both informative and query-relevant texts. The first set of features can identify informative sentences, while the second set of relevant features can aid in the extraction of query-related sentences. These features are related to the subject matter and headlines. However, while Ahuja's derived features can be used for generic summarization, they are insufficient for query-based summarization. However, they have shown that the use of more suiTable features leads to improved summaries generated. They used dataset (DUC)2007.
4. In 2019 Valdes, and et al. [24] proposed a new language-independent solution to the multi-document summarization problem. The authors presented an unsupervised extraction technique that uses a semantic graph to integrate the processing of numerous phrase features with a fuzzy logic perspective. combines the use of a semantic network with the processing of numerous phrase features

using a fuzzy logic perspective.a semantic graph was employed to capture the conceptualization and underlying semantic structure of the textual content, using WordNet database to identify ideas and semantic links between them. The method was tested using the MultiLing 2015 Spanish and English text collections.

5. In 2019 Mallick, and et al [25] suggested a text summarization technique that uses lexical chains and a coherence metric to choose summary phrases.The authors concentrated on leveraging lexical chains to create a summary for a single document. The problem was tackled in three parts: noun phrases of all the sentences are inspected in the first phase, and some of them are chained until no pair of sentences remains for processing. The weights of the individual chains were totaled and sorted in descending order in the second phase.In the final stage, the highest weighted chains were chosen iteratively until the summary length is obtained. It is worth noting that they have used news articles from the BBC Newsfeed.
6. In 2019 Pradeepika and OM [25] suggested a new text summation approach that seeks to locate the best combinations of sentence scoring systems and construct a summary by improving cohesiveness, non-redundancy, and readability. there Many meta-heuristic approaches, such as the Genetic Algorithm (GA), Harmony Search Algorithm (HSA), Cat Swarm Optimization (CSO), and others, have been used to discovered the best weights for scoring systems or relevant sentences for a summary generation. However, for tuning a large number of control parameters, these metaheuristic techniques necessitate significant computational effort. Therefore, the authors adopted the concept of Teaching–Learning Based Optimization (TLBO) in their research, which has a minimum number of control parameters and produces very consistent performance. The key advantage of TLBO is the lack of control parameters. Because TLBO only requires one random number, the variation in output is quite modest in compared to other systems.

7. In 2019, the authors Rezaei, and et al [26] used three distinct regression approaches, including Linear Regression, Decision Tree Regression, and Epsilon-Support Vector Regression (SVR), and proposed explicitly adding document attributes in the feature vector of sentences. In addition, the researchers advocated for the use of features that take into consideration the attributes of a document. They called this feature document-aware, and they used the Pasokh dataset, which contains 100 Persian news documents, each with five summaries. The authors discovered that ROUGE ratings are always relatively high, even for a hastily produced summary. However, they cannot learn to rank by considering dataset sentences as an independent educational instance.
8. In 2020 Zhong, and et al [27] introduced a paradigm shift concerning the way built neural extractive summarization systems by proposing a novel summary-level framework and conceptualize extractive summarization as a semantic text matching problem in which a source document and candidate summaries were extracted from the original text matched in a semantic space. Instead of the commonly used framework of extracting sentences individually and modeling the relationship between sentences. The main objective should be more semantically similar as a whole to the source document than the unqualified summaries. The results of the experiment demonstrated the effectiveness of their method on six benchmark datasets.

This dissertation is different from the studies above because it creates a new developed lexical chain depending on sentences as a whole as a new idea by using universal sentence encoder (USE), cosine similarity, and Develop clustering algorithm proposed. Instead of using wordnet to create a lexical chain depending on words, as is present in the rest of the articles in this field. It is also the creating of the class label and using ROS the method in dataset DUC 2002 that is the basis of this work, as the idea is not found in the section related to researchers' works. Last, reorder location sentences in the candidate summary to form a cohesive summary. Table 1.1 displays a summary of the studies mentioned above.

Table 1. 1 Summary of Previous Studies

No	Authors	Techniques	Type summary	Datasets
1	Saleh, and el al, 2015	Genetic Algorithm (GA).	generic Extractive summary from multi-document	DUC 2002
2	Kadhim, and Saleh, 2016	a Multi-Objective Optimization	generic Extractive summary from multi-document	DUC 2002
3	Afsharizadeh, et al,2018	query-oriented text summarization technique	extracting the most informative sentences	Duc 2007
4	Valdés and et al,2019	fuzzy logic	Offered a new language-independent approach for the solution of the multi-document summarization	MultiLing 2015
5	Mallick, et al.,2019	text summarization technique by constructing lexical chains	Single document	BBC Newsfeed
6	PRADEEPIKA & OM,2019	word2vec, TLBO metaheuristic, and WordNet	extracting multi-document summarization	DUC 2006, and 2007
7	Rezaei, and et al, 2019	approaches, including Linear Regression,	extracting Single-document summarization	Pasokh dataset

		Decision Tree Regression, and Epsilon-Support Vector Regression (SVR)		
8	Zhong, and et al 2020	Neural network	neural extractive summarization systems	CNN/Daily Mail, PubMed, WikiHow, XSum, Multi-News, and Reddit datasets
9	Our proposed system	creates a new lexical chain depending on sentences by using USE, cosine similarity, and Develop clustering algorithm proposed	extracting multi-document summarization	DUC 2002

1.7 Dissertation Organization

The remainder of this dissertation is arranged and organized as follows:

- Chapter Two (Background of The Fundamental Concepts) comprises two parts: the first part gives the basic ideas about Text Mining, and Text Preprocessing. The second part clarifies the basic concepts of clustering, ,USE model, cosine distance, lexical chain word, feature extraction, Multilayer Perceptron Backpropagation Neural Networks (MLPNN) and activation functions, measure evaluation for clustering algorithm, measure evaluation type for text summarization, and USE for data representation.

- Chapter Three (Proposed Methodology to Extract Cohesive Candidate Summary) clarifies the architecture and workflow of the proposed model for topic summarization. It describes this model with the basic architecture in the block diagram, and it provides details of each part.
- Chapter Four (Experimental Results and Discussions) presents the experimental results that are conducted to test the proposed model Afterwards, it evaluates and discusses the performance of the proposed model. Then the results of the proposed cluster algorithm are compared with K-means and hierarchal clustering (HC) to identify which one gives the best result. Also, display results evaluation for candidate summary.
- Chapter Five (Conclusions and Future Work) concludes the overall work presented in this dissertation, summarizes the results and gives some research direction for the future work.

CHAPTER TWO

BACKGROUND OF THE FUNDAMENTAL CONCEPTS

2.1 Introduction

This chapter introduces the key concepts that serve as the foundation for understanding the dissertation work. This chapter discusses Text Mining, Lexical Chain, Cohesion and Lexical Cohesion, Feature Extraction, USE model, Cosine Similarity (CS), and Text Summarization Evaluation Criteria. Then, it goes over clustering, cluster validity, and internal clustering evaluation criteria. It also explains Multilayer Perceptron Backpropagation Neural Networks (MLPNN) and activation functions.

2.2 Text Mining

The area of Text Mining (TM) has gotten a lot of attention in recent years because of the massive volume of text data generated by various sources such as social networks, patient records, health care insurance data, news outlets, and so on [28]. Text data is an example of unstructured data, which is one of the most basic types of information that may be generated in most settings. Humans can quickly comprehend and perceive unstructured text, but robots have a much tougher time understanding it. This collection of writing is, without a doubt, a great source of information and wisdom. As a result, there is a pressing need to develop methods and algorithms for processing this avalanche of text in a range of applications. With some exceptions, text mining procedures are similar to classic data mining and knowledge discovery methods. Knowledge discovery in databases (KDD) is the nontrivial process of obtaining implicitly valid, new, and potentially helpful information from data. Data mining is the process of extracting patterns from data using specific algorithms. While Text mining refers to the extraction of information and patterns that are implicit, previously unknown, and potentially valuable in an automatic or semi-automatic manner from immense unstructured textual data, such as natural-language texts. The goal of KDD is to find hidden patterns and connections in data. As a result, KDD refers to the general process of extracting meaningful information from data, whereas data mining is a single phase in that process. Data can be structured, such as in databases, or unstructured, such as in a plain text file [4, 29].

2.3 Text Preprocessing

Natural Language Processing (NLP) is a branch of computer science that focuses on automated text and language analysis. The preprocessing stage is regarded as one of the most significant processes in text processing [30]. It contains a number of operations, the most notable of which are word tokenization, punctuation removal, sentence tokenization, stop word removal, and so on. It is not necessary to perform all preprocessing methods every time, but the techniques are chosen based on the nature of the task [18].

2.4 Lexical Chain

The semantic distance between the terms is calculated using WordNet to create the lexical chain [31] . WordNet is an online lexical database designed for use under program control. English nouns, verbs, adjectives, and adverbs are organized into sets of synonyms, each representing a lexicalized concept [32]. Words have lexical relationships with one another. WordNet is used to extract these lexical relationships between words [33].

Each word must belong to exactly one chain when lexical chains are computed. However, there are two obstacles: To begin, a word may have more than one meaning (ambiguous word), in which case the proper sense must be identified. Another difficulty is that a word may be related to words in multiple chains. The lexical chain aims to find the best way of grouping the words that will result in the longest and strongest chains.

Among the high-performance characteristics is the lexical chain. As a result, lexical chains can be used to identify the relevance of sentences in a text as an intermediate representation of the lexical cohesiveness that exists across the text. Hence, this feature makes better predictions than many of the other text features. As a result, this feature out predicts several other text features. Even though the weight of the lexical chain feature doesn't seem to be as high as the weights of features like sentence location and sentence centrality, there are other ways to look at how well a text is put together. The sentence centrality and co-occurrence link features support and strengthen the results because they look at how well the text is put together from a different angle [31, 34].

Lexical chaining Consider word sense disambiguation as an example of "semantic techniques," also known as "linguistic approaches," because it attempts to create linkages between words or phrases to provide a partial understanding of the document [35]. Morris and Hirst were the first to put the concept of lexical chaining into practice in 1991. The lexical chain primarily addresses the issue of word sense disambiguation (WSD). Lexical chains are built using the same topic terms as the document. Because several words might relate to the same subject, lexical chains provide a more accurate representation of the subject of a conversation than word frequency. Even if these procedures do not divide senses, they can generate concepts [17].

2.5 Cohesion and Lexical Cohesion

Cohesion is described as the connectedness of a text's exterior elements. Cohesion happens when the interpretation of one part of the text is dependent on other elements to the point that one element cannot be understood without the other elements. Whereas coherence differs from cohesion, coherence can be described as the relationship between words or phrases in a text such that it is one part of the discourse utilized to explain a fact or event. As a result, coherence is tied to the paragraph's interrelation and interfaith order. Cohesion (unity) and coherence are concepts used in the speech, both in paragraphs and entire works. The coherence and cohesion of the discourse become measurements of the readability of the discourse's contents [36, 37]. To learn more about coherence, it is classified into two categories: grammatical and lexical [36]:

1.The forms of cohesion in the grammatical aspect can be described as follows [38]:

- **Reference** is the heart of the special information that has been marked to be found again. It is in the form of referential meaning, which is the name of the thing being talked about. There are words and phrases or other grammatical units that show that there is a reference. In other words, they talk about the use of pronouns for things. This is how it works: "Dr. Kenny lives in London." He's a doctor. "He" in the second sentence refers to "Dr. Kenny" in the first one.

  Substitution The placement or replacement of anything with other elements is referred to as substitution. Substitution takes the form of specific language elements that replace the language elements that come before or after them. In other words, it employs an indefinite article to refer to a noun. The term "one" corresponds to the phrase "vanilla ice cream cone" in the example, "As soon as John was given a vanilla ice cream cone, Mary wanted one too.".

- **Ellipsis** The removal of an element. Imposing occurs when some crucial structural pieces are removed and can only be retrieved by referencing a previous part in the text. This ellipsis is cohesiveness in the form of the removal of previously specified constituents. In other words, it implies a noun but does not repeat it. says the statement "Do you have a pencil? 'No, I don't”, the word "pencil" is implied but not explicitly stated in the second phrase. Lexical cohesion is the most definite and easiest to find of these cohesion patterns.

- **Conjunction** When two or more ideas in a speech are linked together, this is called "conjunction." Conjunctions express certain meanings that show the preconditions for the presence of other components in the discourse.

2.While The forms of cohesion in the lexical aspect are described as follows [36]:

- **Repetition** is the repetition of lingual units (sounds, syllables, words, or sections of phrases) that are thought to be crucial to applying pressure in a proper context.
- **Synonyms** are alternative names for the same item or expressions that have similar meanings to other terms. Synonyms help to develop meaningful linkages between particular words and other terms in a conversation.
- **Collocation** entails particular correlations in the use of word choices that are frequently employed together. Collected words are terms that are often used in a specific domain or network.

Typically, text linguistics attempts to determine the structure of a text by analyzing it in terms of the two most significant criteria: cohesion and coherence. In a text, cohesion refers to the use of other elements to explain or interpret an aspect within the text, as well as giving explanations based on these elements. It is worth noting that tools of cohesion can be described as semantic linkages between one element and another that play a significant role in understanding the first element within the text [39].

Finally, lexical cohesiveness has been used to discriminate between significant and irrelevant sentences in the text. This is thought to be beneficial. Lexical cohesiveness divides the text into segments, each of which consists of a series of sentences on a given topic. The essential subjects are included in the most relevant segments. The sentences chosen to convey the summary are from the more relevant segments. [40].

2.6 Feature Extraction

The extraction of features is crucial in the field of topic summarizing and is dependent on the goal of the text summary. The purpose of feature extraction is to find the features in each sentence. As a result, the important sentences from numerous documents can be identified and chosen to construct the candidate's summary. This section will clarify the implementation procedures for some of the features used in this dissertation. The following are some of the features:

1. **Centrality**: This feature falls under the category of informative score features. The sentence's centrality implies that it is similar to other sentences. A document (or a collection of documents) is represented as a matrix, with nodes representing sentences and links connecting them weighted according to their similarity. The centrality of a node can be determined by computing its degree or by running a ranking algorithm. After calculating the centrality score for each sentence, the sentences are sorted in reverse (i.e. descending) order, with the highest-ranking ones included in the summary. If a sentence has a greater centrality degree, it is the best contender for inclusion in the summary, and its score is calculated as follows in equation 2.1 [41]:

$$\mathbf{Centrality(S_i) = \sum_{j=0}^{n} CosSim(S_i\,, S_{(n-j)})}\text{...............}\ \mathbf{2.1}$$

Where S_i represent sentence and CosSim is mean cosine similarity distance.

2. **Proper noun (Number noun word feature (NNWF))**: This characteristic is categorized as a readability feature; sentences with the most noun words are considered to be significant and are more likely to be included in the candidate summary. Generally, noun terms refer to individuals, actions, cities, and objects. The proper noun is calculated as equation 2. 2 by dividing the total number of noun terms in the sentence by the sentence length [23]:

$$\mathbf{NNWF} = \frac{\sum \mathbf{noun\ words\ (S_i)}}{\mathbf{len\ (S_i)}} \quad \mathbf{.............\ 2.\ 2}$$

Where S_i represent sentence.

3.**Sentence Position Feature (SPF)**: In general, the first and last sentences in a text are considered the most important [23], but in the study, the golden summary existence in the dataset used in this dissertation found that there were also important sentences in the middle of the text. This is what is referred to as the Sentence Position Feature (SPF) in the informative score features. Equation 2. 3 was used to calculate this characteristic [42].

$$\mathbf{SPF} = \mathbf{Max}\left(\frac{\mathbf{1}}{\mathbf{i}}, \frac{\mathbf{1}}{\mathbf{N-i+1}}\right) \mathbf{...............\ 2.\ 3}$$

Where i is the index sentence, and N is the number of sentences in the document. In this dissertation, this feature is used to extract the important sentence position from each document even if it is in the middle of the text as mentioned above. Because this work deals with multiple documents.

4. **Sentence Length Feature (SLF)**: This feature is categorized as a readability feature because of the length of the sentences it contains. It is critical to assign importance to a sentence based on the length of the sentence to the rest of the phrase since big sentences include more information than short sentences. This feature is computed by dividing the length of each sentence in the document by the length of the longest sentence in the same document as equation 2. 4 [23].

$$\mathbf{SLF(S_I)} = \frac{\mathbf{Length(S_I)}}{\mathbf{The\ Longest\ Sentence\ In\ The\ Document}} \quad \mathbf{...........2.\ 4}$$

This work has taken advantage of this feature to calculate the length of the sentence so that the sentence that achieved a higher score is considered a long sentence.

5. **Headline Feature (HF)**: This feature is classified within readability features, the similarity of each sentence in the document with the title of the document is calculated as semantically using cosine similarity. The sentences that achieve a high degree in this characteristic mean that they contain special important words related to the concept of texts. This feature can be calculated through equation 2. 5 [42]:

$$\mathbf{HF(S_i)} = \frac{\sum_{i=0}^{n} S_i \times H}{\sqrt{\sum_{i=0}^{n} (S_i)^2} \times \sqrt{\sum_{i=0}^{n} (H)^2}} \quad \mathbf{\ldots\ldots\ldots\ldots\ldots.\ 2.5}$$

Where H represents the Headline sentence in the document and S_i represents each sentence in the document.

6.**Sentence Numerical Feature (SNF)**: This feature is classified within informative score features. It searches for sentences that contain numbers and gives weight to that sentence based on equation 2. 6 [23]:

$$\mathbf{SNF(S_i)} = \frac{\mathbf{CNW(S_i)}}{\mathbf{len(S_i)}} \quad \mathbf{\ldots\ldots\ldots.\ 2.6}$$

Where $\mathrm{CNW(S_i)}$ implies count numerical words existence in a sentence S_i, and $\mathrm{Len(S_i)}$ is represent sentence length.

7. **Positive Feature (PF):** This feature is classified within informative score features. It works to find the positive words in each sentence and give weight to them. Usually, these words are like ”defined as”, ”called”,”significant”, ”means”, ”important” increasing sentence importance more often. Can be calculated by equation 2.7 [42]:

$$\boldsymbol{PF(S_i)} = \frac{\boldsymbol{CPW}}{\boldsymbol{len(S_i)}} \quad \mathbf{\ldots\ldots\ldots\ldots\ldots 2.7}$$

Where CPW represents the count of positive words existing in a sentence S_i.

8. **Negative Feature (NF):** This feature is classified within informative score features. It works to find the negative words in each sentence and give weight to them. Usually, these words are like ”not”, ”but”,” unless”, ”although”,” however” less sentence importance more often. Can be calculated by equation 2.8 [42] :

$$\boldsymbol{NF(S_i) = \frac{CN}{len(S_i)}} \quad \textbf{\dots\dots\dots\dots\dots 2.8}$$

Where CN represents the count of negative words existing in a sentence S_i, and $len(S_i)$ represents sentence length.

2.7 Universal Sentence Encoder (USE) Model

Typically, sentence embeddings (which are commonly constructed on top of word embeddings), and encoding grammatical structures that are not present in the task-specific training data can be employed to further boost generalization capabilities. As an outcome, high-quality universal sentence representations are extremely desirable for a wide range of downstream NLP applications. These representations, or embeddings, differ in that they are not always trained to perform well on a single task. Additionally, their ability to acquire data that may be used in any type of system or pipeline, on a wide range of tasks [43]. one of the models used for sentence embeddings is the Universal Sentence Encoder (USE) produced [44]. These models used pre-trained sentence embedding which explained that it robust transfer task performance rather than models which used pre-trained word embedding like the word 2 vector model which was produced by [45], and the Glove model which was produced by [46]. Because word embeddings are especially effective in cases where there is limited training data, leading to sparsity and poor vocabulary coverage, which in turn lead to poor generalization capabilities [43, 47]. This dissertation used a Universal Sentence Encoder (USE) model for constructing sentence embeddings.

in 2018, Daniel Cer, and et al produced the USE model. They discovered that transfer learning with sentence embeddings outperforms word-level transfer. Furthermore, for a transfer task, they reported very strong performance with small quantities of supervised training data using transfer learning via sentence embeddings. Thus, USE model showed good transfer to a variety of different NLP tasks in sentence embedding. This model includes two models proposed based on trade-offs in accuracy against inference speed. The first is the transformer encoder model which builds sentence embeddings using the transformer architecture's encoding sub-graph. This sub-graph employs attention to calculate context-aware representations of words in a sentence that accounts for both the ordering and identity of all other words. The transformer encoder has better accuracy on downstream tasks but higher memory and computes resource usage due to complex architecture. The Second encoding strategy employs a Deep Averaging Network (DAN) model, which averages input embeddings for words and bi-grams before passing them through a feedforward Deep Neural Network (DNN) to produce sentence embeddings. The DAN encoder, like the transformer encoder, takes a lower-cased tokenized string as input and generates a 512-dimensional sentence embedding [44]. The Deep Averaging Network (DAN) encoder will be discussed in detail because this architecture has been used in this work.

- **Deep Averaging Network (DAN) encoder**

This encoder is based on the architecture proposed by Iyyer, and et al. [48]. The authors Iyyer, and et al. summarized how the deep averaging network (DAN) works in three simple steps:

Step1: Compute the vector average of the embeddings associated with tokens sequence input.

Step2: Take the average and run it through one or more feedforward layers.

Step3: Apply (linear) classification on the representation of the final layer.

Usually, Deep feed-forward neural networks are designed with the idea that each layer learns a more abstract representation of the input than the preceding one [49]. DAN applied on Neural Bag-of-Words (NBOW) model. The NBOW simplicity, consider text classification: convert an input sequence of tokens X to one of k labels as equation 2. 9 proposed by [48]:

$$z = g(w \in X) = \frac{1}{|X|}\sum_{w \in X} v_w \text{ ……2.9}$$

Where z is the vector representation for input text X, by averaging the word vectors $v_{w\in X}$,g represents composition function, v_w represents a sequence of word embeddings, X represents an input sequence of tokens. Feeding z to a softmax layer induces estimated probabilities for each output label as Equation 2. 10 [48]:

$$\hat{y} = softmax(W_s \times z + b) \text{………….. 2.10}$$

where W_s is $k \times d$ matrix for a dataset with k output labels, and b is a bias term, and Softmax activation function as equation 2. 11 [50]

$$Softmax\ (q) = \frac{\exp{(q)}}{\sum_{j=1}^{k} exp\ (q_j)} \text{………2.11}$$

They abled further transform z by adding more layers before applying the softmax. The authors have supposed to have n layers, $z1 \dots . n$. They computed each layer as equation 2. 12 [48].

$$z_i = g(z_{i-1}) = f(W_i \times z_{i-1} + b_i) \text{…….2.12}$$

In the beginning, DAN works embeddings for words, and bi-grams present in a sentence are averaged together. Then, they are passed through a n-layer feed-forward deep DNN to get 512-dimensional sentence embedding as output. The embeddings for word and bi-grams are learned during training as showed the DAN layout in Figure 2.1 by sentence1

“Predator is a masterpiece” [48, 44]. The sentence2 “Hello world” is example explain work DAN as Figure 2.2 [51].

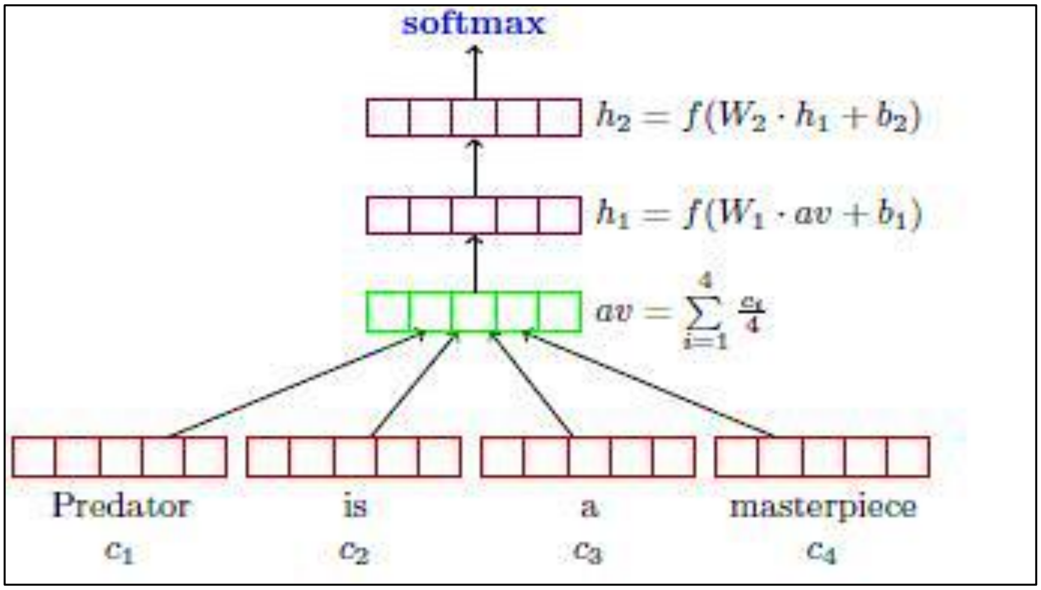
softmax
$h_2 = f(W_2 \cdot h_1 + b_2)$
$h_1 = f(W_1 \cdot av + b_1)$
$av = \sum_{i=1}^{4} \frac{c_i}{4}$
Predator
c_1
is
c_2
a
c_3
masterpiece
c_4

Figure 2. 1 Layout DAN which Applied on Sentence1 [48]

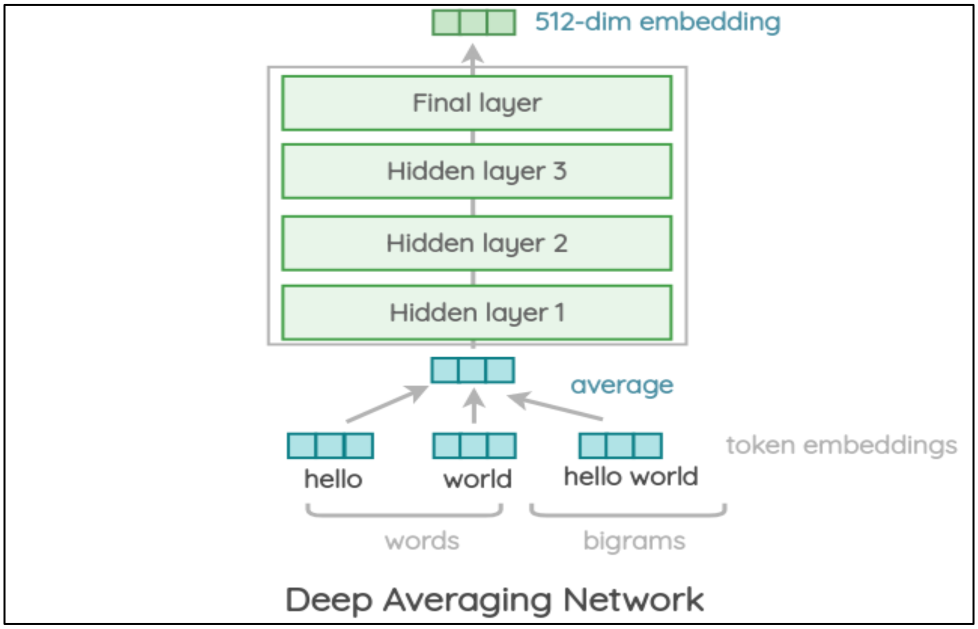

Figure 2. 2: Layout DAN which Applied on Sentence 2 [51]

2.8 Cosine Similarity (CS) Metric

Many metrics, such as the Euclidean distance-based metric, Cosine, Jaccard, Dice, and Jensen-Shannon Divergence-based metric, have been presented in recent years to cope with various types of information retrieval and natural language processing difficulties. One of the most commonly used metrics is the angle among two vectors, which is the cosine because cosine similarity is used in high dimensional spaces [52] . The multiplication of two normalized vectors is used to calculate it. The cosine similarity between two N dimension vectors V and C is determined using equation 2. 13 [53]:

$$\textbf{Cosine Distance}(\boldsymbol{V},\boldsymbol{C}) = \frac{\sum_{i=0}^{n} \boldsymbol{V_i} \times \boldsymbol{C_i}}{\sqrt{\sum_{i=0}^{n} \boldsymbol{V_i^2}} \times \sqrt{\sum_{i=0}^{n} \boldsymbol{C_i^2}}} \dots\dots\dots\dots\dots\dots \textbf{2. 13}$$

Note, this dissertation has been used N dimension which represent 512 the fixed length dimension because sentences have been converted into embedding vectors by using Universal Sentence Encoder (USE) model.

Cosine similarity is a metric that is often used to execute tasks such as information retrieval and data mining, particularly in high-dimensional positive spaces. Unlike the Euclidean distance, which is sensitive to even minor deformations, cosine similarity is more concerned with directions [54]. Cosine similarity is the most common inner product family measure of the degree of similarity between two vectors. It can be used to represent the degree of similarity between two sentences within the context of text classification [55].

2.9 Evaluation Criteria for Text Summarization

The evaluation of the performance of all summarizing jobs is a difficult problem in and of itself. Recall − Oriented Understudy for Gisting Evaluation ($ROUGE$) is the most often utilized evaluation technique nowadays [56]. It provides measurements for automatically judging the quality of summaries by contrasting them with ideal human-created summaries (such as the count of overlapping n-grams or word sequences). However, because these measurements do not include the meaning of words, synonyms, for example, are not considered equal. Because automatic summarizing is evaluated by comparing the system summary to a human-made one [57]. ROUGE scoring expects system summaries to recover the contents of reference summaries exactly. ROUGE scores stay low unless the system summary is fully similar to the reference summary [58].

ROUGE is a recall-based quantitative analysis metric. It computes the number of overlapping n-grams between the computer-generated summary and the reference summaries written by humans. Previous research has found that ROUGE scores are highly linked with human judgments of summary quality. There are many variants of the ROUGE metric, for example, 1) The bigram overlap between the candidate computer-generated summary and the reference summaries is measured by ROUGE-2. 2) ROUGE-N calculates the n-gram overlap, which includes ROUGE-1. 3) For contrast, ROUGE-L measures the overlap in the Longest Common Subsequence. 4) ROUGE-S calculates over-laps in skip-bigrams or bigrams with arbitrary gaps. 5) ROUGE-SU4 compares skip-bigrams to unigrams. Some researchers have used human evaluation along with the ROUGE Score to find the overall quality of the summary [59] [60]. The Rouge scale degree might be useful in covering. However, it is ineffective for determining coherence and other aspects such as non-redundancy because these metrics only compute the repeatability of N-grams. [59]. Can compute the Precision, Recall, and F-measure for any of these metrics in the following [61]:

A. ROUGE recall refers that how many words of the candidate summary are extracted from the reference summary. The equation 2. 14 is used to calculate recall:

$$\boldsymbol{Recall} = \frac{\boldsymbol{number\ of\ the\ overlaping\ words}}{\boldsymbol{Total\ words\ in\ golden\ summary}} \quad \ldots\ldots 2.\,9$$

B. ROUGE precision refers to how many candidate summary words are relevant. The equation 2. 15 is used to calculate precision:

$$\mathbf{Precision} = \frac{\boldsymbol{number\ of\ overlaping\ words}}{\boldsymbol{Total\ words\ in\ candidate\ summary}} \text{ ....2. 10}$$

C. F-measure provides the complete information that recall and precision provide separately. equation 2. 16 to calculate **F-measure:**

$$\mathrm{F-measure} = \frac{(1+\beta^2)\ R \times P}{R+\beta^2 \times P} \text{ ....2. 11}$$

Where β=1, R represents recall, and P represents precision.

2.10 Clustering

Clustering is a typical approach for detecting and analyzing structure in data. In other terms, it is the process of generating groupings (clusters) of similar patterns from a given collection of input data. Clustering is being used in numerous disciplines, including web mining and text mining, business and marketing, machine learning, pattern recognition, image analysis and segmentation, information retrieval, and bioinformatics [62]. Clustering is an unsupervised aggregation approach. This technique has a wide range of applications in domains such as medicine, business, imaging, marketing, image segmentation, chemistry, robotics, and climatology. It is a branch of data mining technology that is particularly efficient in selecting useful information from a dataset [63].

Cluster analysis methods are classified as statistics and informatics. One of the cluster constraints is that two items from the very same cluster are more similar than two objects from different clusters, and the partitioning procedure should achieve two crucial characteristics. The first characteristic is cluster homogeneity (data belonging to the same cluster should be similar). The second characteristic is cluster heterogeneity (data from two or more different clusters should be dissimilar) [64, 65].

Most clustering methods involve, implicitly or explicitly, a similarity metric among patterns, the selection of which is challenging if no prior knowledge about cluster forms or structure is available. The Euclidean distance is the most commonly used pairwise distance between patterns in most clustering techniques. However, numerous additional measurements, such as the Mahalanobis distance, can be utilized [66].

There are two main categories of clustering methods including partitional [67, 68, 69],Hierarchical clustering [67, 68, 69] , Grid-based methods [68], and Density-based methods [68, 69] . The following subsections explain these clustering methods in some details**.**

2.10.1 Partitional Clustering

Partitional clustering algorithms aim to partition the input data into a fixed number of clusters [70]. The partitional methods seek a partition that improves a sufficiency criterion function locally. Typically, the user must specify the number of clusters ahead of time. Selection a right number of clusters in a model considered as a problem itself, which is frequently non-trivial, especially for real-world datasets [66].

Partitional clustering algorithms have two primary drawbacks. In the beginning, the quality of data clusters cannot be guaranteed because partitional clustering relies significantly on predetermined criteria, such as the number of clusters. Second, partitional methodologies discover all clusters as a partition of the data at the same time and do not impose a hierarchical structure [71].

There are two types of partitional clustering: hard clustering and fuzzy [72]. Hard clustering produces a division where each item in the dataset is assigned to a single cluster. The well-known k-means algorithm is a member of the partitional clustering algorithm family [70]. This algorithm is starting with a random set of centroids, then assigns each item to the centroid closest to it. Then, for each group, a new central point is determined depending on its members, and pattern assignments to their closest centroids are altered as needed. The algorithm completes when no pattern reassignments are required or after a predetermined amount of time has passed [62]. Clusters, on the other hand, can overlap in numerous real-world problems, resulting in many things having features from multiple clusters [73]. Fuzzy clustering generates a fuzzy division that indicates the membership degree to which each pattern belongs to the specified cluster. Due to the fact that fuzzy clustering is more natural than hard clustering, items on the boundaries of many clusters are not required to fully belong to one of the clusters; rather, they are awarded membership degrees ranging from 0 to 1 to indicate their partial connections [70]. The Fuzzy C-Means (FCM) clustering technique is one of the fuzzy clustering methods that are available. The standard FCM clustering algorithm is a non-hierarchical clustering method. Its goal is to provide a fuzzy partition of a set of patterns in c clusters and a corresponding set of prototypes in such a way that a criterion measuring the fitting between the clusters and their prototypes is locally minimized [74].

2.10.2 Hierarchical Clustering (HC)

A set of patterns is sorted into hierarchical clusters and shown as a tree called a dendrogram. A dendrogram is a tree with each node representing the integration of two or more partitions [62]. As a result, it may categorize and divide a dataset into numerous levels using distance or density functions. BIRCH (balanced iterative reduction and clustering utilizing hierarchies) [75] is an example of a clustering technique that builds clusters rapidly and effectively using a tree structure. Chameleon finds data clusters by evaluating data similarities and grouping them together. Clustering using Representatives (CURE) [76] is a technique that can be used to find non-spherical cluster structures in a large database [68].

The qualities of hierarchical representation are advantageous for the viewing and interpretation of clustering findings. The hierarchical clustering technique groups items using a nested sequence of partitions via an agglomerative (bottom-up) or divisive (top-down) strategy [77].

In agglomerative clustering, each unique pattern is first assigned to a cluster that contains only that pattern. Then, the two clusters that really are nearest to each other are integrated into a new group, and so on until one gets a cluster that contains all of the patterns. [62, 77].

Divisive clustering typically works in the opposite direction. It starts with a single cluster containing all of the patterns and divides them until a stopping requirement is fulfilled (usually upon obtaining a partition of singleton clusters) [70]**.**

Hierarchical clustering has two major advantages. To begin, the number of clusters does not need to be stated in advance. Second, they are unaffected by the beginning conditions. The main disadvantage of hierarchical methods is that they are static, which means that data points allocated to one cluster cannot be moved to another. Furthermore, because to a lack of knowledge regarding the global structure or size of the clusters, they may fail to distinguish overlapped clusters [78]. Additionally, HC faces a fundamental difficulty in data analysis, where provided data points and their pairwise similarities are represented as a tree, with leaves representing data points and internal nodes representing clusters. Despite the availability of HC algorithms. But the HC theory is regarded as undeveloped due to the lack of a "global" purpose [79].

2.10.3 Density-based methods

Density-based clustering (DBCL) is based on the idea of determining a region's density. The goal of DBCL is to discover clusters with appropriate noise filtering at various levels of granularity. The DBCL's density idea allows for the separation of compact regions in the data space from noise. Clusters are detected in DBCL as places with a higher density than the rest of the data space These methods are designed to find clusters of densities that are reasonably uniform across the data space [69]. DBCL have the ability to discover different shapes of clusters. For example, Density Based Spatial Clustering of Applications with Noise (DBSCAN) [80]. DBSCAN aid in the detection of arbitrary-shaped clusters. DBSCAN is sensitive to input parameters, however, ordering points to identify the clustering structure (OPTICS) [81].

As a result, density-based approaches can be used to uncover distinct cluster shapes. For example, density-based spatial clustering of applications with noise (DBSCAN) is a well-known algorithm for identifying the arbitrary shape of clusters, and many individuals have offered improved approaches to overcome any shortcomings and increase efficiency. The input parameters of DBSCAN are very important. However, using ordering points to determine the clustering structure (OPTICS) may be able to avoid this problem which from impacting the clustering results. It, on the other hand, is unable to obtain reliable cluster results. Density-based clustering (DBCL) is a method that uses a kernel density estimation model to identify the high density of clusters of any shape [68].

2.10.4 Grid-based methods

Grid-based approaches create the dataset is stored using a grid layout, with each grid acting as the fundamental unit of a cluster. Grid-based clustering is especially useful when dealing with large datasets [82]. The main idea is to create a spatial summary of the data by covering the data space with a grid. Each non-empty cell of the grid is weighted by the number of original data points it contains, as shown in Figure 2.3 Clustering is accomplished by combining adjacent dense cells to form clusters [83].

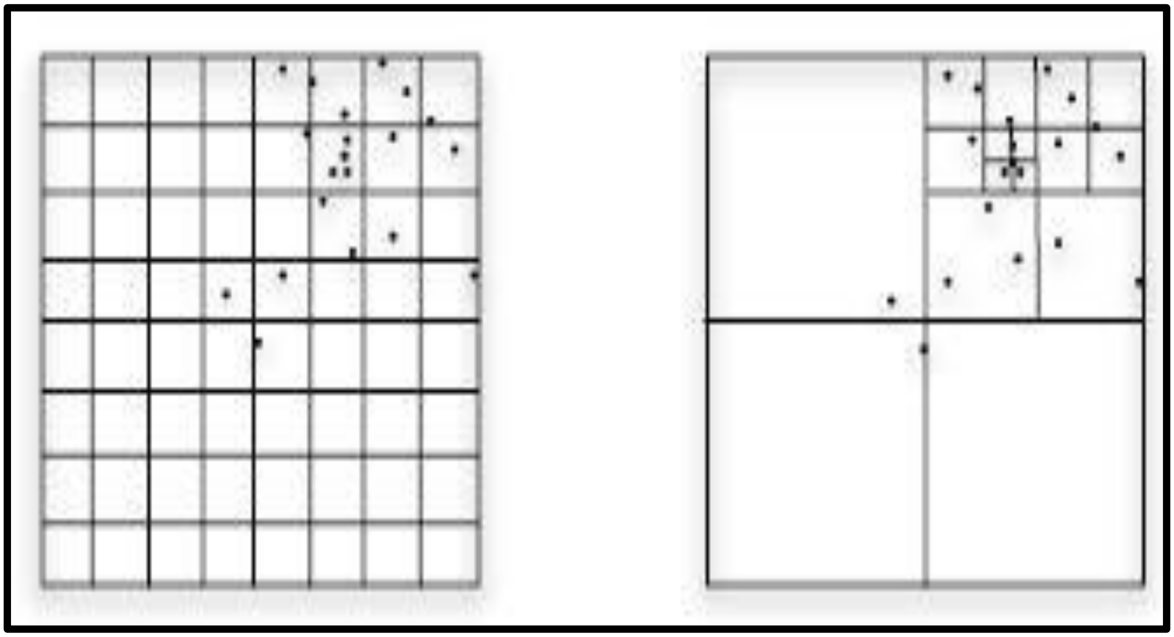

Figure 2. 3 Summaries of Datasets. Left A Regular Rectangular Grid. Right An Adaptive Hypercubic Grid.

The grid-based clustering method thereby divides the data space into fragmented grid cells, with clustering focused on the space surrounding the data points rather than the data points themselves. When seeking to locate clusters of observations that may be displayed using grids, it is crucial to understand how near or far apart the observations are from one another. The distance between individual observations or clusters of observations may be understood in the Euclidean sense or in an analogous manner. Directly calculating proximity by examining each individual data point, or indirectly by selecting a representative point within a grid cell, is possible. The distance average has always been employed and regarded as an exact quantitative measure of the proximity of observational groups [84].

Usually, Grid-based approaches are comparable to density-based clustering, except instead of individual points, local densities and neighborhood interactions occur between cells [83] . Furthermore, computing grid distances using the distance average is time-consuming and expensive, and thus cannot be used to fully exploit the benefits of grid-based clustering [84] .

2.11 Cluster validity

Clustering validity is an important criterion for assessing the clustering's quality. As a result, cluster validation is required to verify the goodness of clustering following partition. Many statistical and numerical measurements attempt to determine how effective the clustering algorithms and starting parameters were for the dataset in question. The cluster validity index is a numerical measure that can be used to represent these measures. Internal, external, and relative criteria are the three types of cluster validity indexes [85, 86]. Internal criteria assess the quality of clusters for data by extracting information from data and clusters alone, such as compactness of data points within a cluster and cluster separation, and include several measures such as the Davies Bouldin Index (DBI), Dunn Index, Silhouette Coefficient, the sum of squared error, scatter criteria, trace criteria, and determinant criteria. External criteria, on the other hand, assess results based on a preset structure for data that is provided in the form of labeling in addition to clustered data. The major goal of this method is to find a statistical measure for the similarity or dissimilarity between generated clusters and labels [87], which includes measures like Variation of Information, Jaccard Coefficient, and Fowlkes-Mallows Measure. Finally, relative criteria are used to compare two clustering algorithms with the same data and differing beginning parameters such as the number of clusters. While most approaches utilize internal indices to evaluate relative, other research employs external criteria [85].

This section will go over one of the internal criteria, the Davies Bouldin Index (DBI), in detail because it was used to assess the quality of the clustering algorithms employed in this dissertation.

2.11.1 DBI clustering evaluation

The Davies Bouldin Index (DBI) is used to analyze linkages between clusters and to evaluate correlation content within each cluster. Finally, it assigns a grade. This score is good whenever it has a positive and low number. This implies that this strategy is strong. The DBI adds a scattering measure SC_i to measure scattering inside the same cluster and maximizes the scattering measure to cluster center isolation ratio. Let C. deliver the DBI for several clusters. As a result, the DBI takes into account the average case of each cluster by using the mean of each cluster. the equation for scattering measure can be written as equation 2. 17, the equation for measure of how good the clustering scheme can be written as equation 2. 18,and the equation for DBI can be written as equation 2. 19 [88, 63].

$$\boldsymbol{SC_{i,q} = \left(\frac{1}{|A_i|}\sum_{x_j \in A_i} \left\| x_j - v_i \right\|^q \right)^{\frac{1}{q}}} \quad \ldots\ldots \quad \mathbf{2.12}$$

$$\boldsymbol{R_i = max_{j=C\&j\neq i}\frac{SC_{i,q}+SC_{j,q}}{\|v_i - v_j\|}} \quad \ldots\ldots. \quad \mathbf{2.18}$$

$$\therefore \boldsymbol{DBI = \frac{1}{C}\sum_{i=1}^{c} R_i} \quad \ldots\ldots. \quad \mathbf{2.19}$$

Where SC_i represents scattering measure, A_i represents the size of cluster i , x_j be an *n*-dimensional feature vector allocated to the cluster, v_i represents the centroid of the cluster, R_i represents a measure of how good the clustering scheme; and C represents a number cluster.

2.12 Overview of Random Oversampling (ROS) method

The classification problem for imbalanced datasets has been a prominent research subject in recent years. Unbalanced data is one of the top ten most difficult problems in data mining research. As a result, the classification challenge for imbalanced datasets is prevalent in many data mining disciplines. Almost all algorithms suffer from the problem of an imbalanced dataset, in which some classes have more occurrences than others, causing biases in classification and poor generalization performance. Real-world applications include fraud detection, bioinformatics, text classification, the medical profession, and so on. The minority is the most intriguing in these applications, and identifying it is critical. This necessitates a fairly high detection rate for the minority class while allowing for a small error rate in the majority class because the cost of misclassifying a majority instance might be quite low. The issue of class disparity is critical. It has the potential to significantly reduce the performance achievable by typical learning methods that presume a balanced class distribution [89].

ROS is the simplest technique to balance the imbalanced nature of the dataset. It is an effective way to solve this problem at the data processing level so that it balances the data by replicating the minority class samples. This does not cause any loss of information, but the dataset is prone to overfitting as the same information is copied which reduces the classification performance of the model on invisible data [90].

2.13 Artificial Neural Network (ANN)

It is one of the most extensively used modeling tools in reservoir characterization, and it is known as the Artificial Neural Network (ANN). ANN has the strength of getting a better modeling capability by combining with other soft computing techniques. When used with other soft computing techniques (called hybrid models) like fuzzy logic, genetic algorithm, the limitations of ANN like overfitting, parameter selection, can be overcome [91].

2.13.1 The Structure ANN

An Artificial Neural Network (ANN) is a collection of highly interconnected processing components known as neurons, which are analogous to biological neurons. It has been used to construct intelligent systems for prediction, function approximation, pattern recognition, feature learning, optimization, and control, among other things. A multilayer perceptron (MLP), often known as a Feed-Forward Neural Network (FFNN), is the most fundamental and simplest artificial neural network model. Each neuron represents a mapping, especially with multiple inputs and a single output. The output of the neuron is a function of the sum of its inputs. The function used in the output of a neuron is called an activation function. The single output of the neuron can be applied as an input to some other neurons, and therefore the symbol for a single neuron shows the number of arrows coming from the neuron. Signals that are sent from input to output are weighted sums indicating the direct relationship between the two layers. The addition of layers between input and output makes the neural network contains many layers called

Multi-Layer Perceptron (MLP) [92]. The MLP's nodes are linked in such a way that information can only flow forward. Each succeeding layer receives input from the preceding layer, and the process is repeated. In particular, the activation function of the hidden layer is in charge of the nonlinear transformation of the input characteristics into a feature that is easier to represent and more important for choosing the destination. The sigmoid function is the most often encountered activation function. As depicted in Figure 2.4, the architecture of the multilayer perceptron can be seen [93].

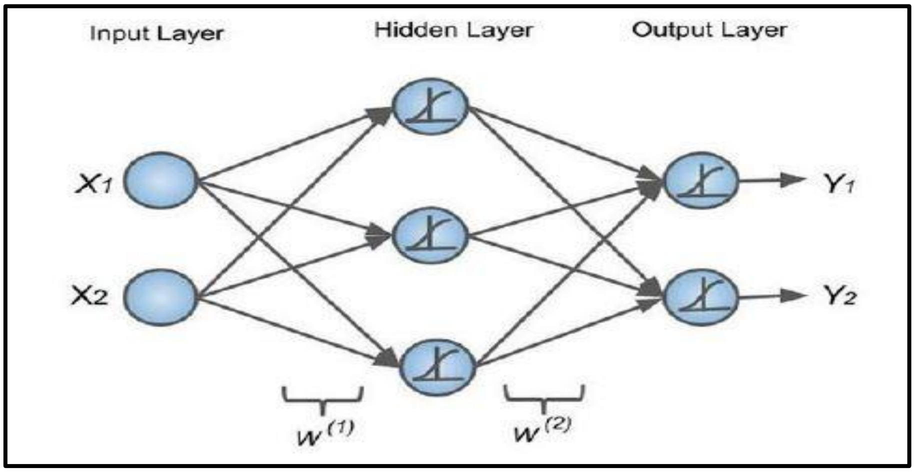

Figure 2. 4 Multilayer Perceptron Architecture X= [X_1, X_2] Is The Input Vector, Y= [Y_1, Y_2] Is The Output Vector, And W=[$W^{(1)}$, $W^{(2)}$] Is The Vector Of Weight Matrices For Layer 1 And Layer 2 [93].

Any ANN platform's utility is dependent on its capacity to train the network utilizing methods such as error backpropagation, which is a critical need for utility. A large amount of computing time and resources are required for such training, and it is normally ideal for error backpropagation to be performed on the same platform as the training process [94]. It is possible to circumvent the constraints of ANN, such as overfitting and parameter selection. Because of its capacity to simulate complex nonlinear functions, artificial neural networks are a widely popular machine learning technique. neural network with one hidden layer is adequate for approximating any continuous function with an adequate number of hidden neurons and any nonlinear function with an adequate number of hidden neurons. However, when dealing with complicated issues such as time series, computer vision, and speech recognition, increasing the number of hidden layers may be beneficial in terms of modeling capabilities and efficiency [94, 93].

2.13.2 Backpropagation algorithm

The backpropagation algorithm, which was first presented in the 1970s, is the most extensively used ANN learned skill. The algorithm gained popularity after 1986, and for the first time, the perceptron learning technique could be used for multilayer perceptron's. The overall training procedure for ANNs employing backpropagation is separated into two phases: forward propagation and backward propagation. In forwarding propagation, data is transmitted to the output layer via hidden layers coupled by random weights. The model's anticipated output at the end of the forward pass may differ from the desired output. The network's weights are modified to get the projected output as near to the intended output as possible. The weight is adjusted during the backward propagation step. The derivative of the mistake (the difference between the intended and projected output) is calculated and propagated backward in backward propagation. This is referred to as "backpropagation via gradient descent" in technical terms. The error derivative is used to change the weights so that the output error is as small as possible. The backpropagation algorithm in its entirety can be summarized as follows [93]:

1- Initialize the weights of the network to small random weights.
2- Present an input vector from the training dataset to the network.
3- Propagate the input to generate the output which is called the Feed Forward phase of the input.
4- Calculate the error by comparing the estimated output of the network with the desired output.
5- Propagate the error backward through the network which is called backpropagation of error.
6- Adjust the weight in such a way that it minimizes the error.
7- Repeat the above steps 2-6 until the error doesn’t improve anymore.

The following are some of the benefits of backpropagation: 1) Backpropagation is a fast, basic, and straightforward technique that is simple to program. 2) Aside from the number of inputs, there are no parameters to tweak in this model. In addition, because it does not require prior knowledge of the network, it is a flexible method. 4) It is a standard method that, in most cases, is effective. 5) It is not necessary to provide any specific description of the characteristics of the job to be learned [95].

2.13.3 Activation Functions

The Activation Function (AF) is critical in the training of neural networks. They provide the model with the necessary non-linearity to learn complex representations. Piecewise linear functions and locally quadratic functions are the two basic types of activation functions. A piecewise linear function is made up of a small number of linear segments that are defined over an equal number of intervals, which are usually of the same size. A piecewise linear function can be used to represent the Rectified Linear Unit (ReLU) variation. The absence of curvature in every interval indicated by the breakpoint is a distinguishing feature of these sorts of activation functions. While a locally quadratic function is a non-linear smooth activation function with a nonzero second derivative. the Sigmoid, Hyperbolic Tangent, Rectified Linear Unit (ReLU), Parametric Leaky Version of a Rectified Linear Unit (PRelu), Exponential Linear Unit (ELU), Scaled Exponential Linear Unit (SELU), Mish, and softmax are some examples of activation functions [96] [97].

The sigmoid activation function, which is used in this dissertation, will be explained in-depth in the following.

- **Sigmoid Activation Function**

  The step activation function's fundamental flaw is that it is non-differentiable. As a result, it can't be utilized to calculate backpropagation neural network error coefficients. Instead, the Sigmoid Activation Function (SAF) is used. [51]. The Sigmoid is a nonlinear AF that is mostly utilized in feedforward neural networks. It's a bounded differentiable real function with positive derivatives everywhere and considerable smoothness, defined for real input values. In deep learning (DL) architectures, the sigmoid function appears in the output layers. They've been utilized successfully in binary classification challenges, modeling logistic regression tasks, and other neural network areas for forecasting probability-based output. It has a smooth derivative and is non-linear by nature, as seen in Figure 2.5. The Sigmoid activation function is also known as the logistic function or the squashing function in some publications. The Sigmoid function study resulted in three sigmoid AF versions that are used in DL applications. Equation 2. 20 can be used to compute the Sigmoid activation function [95] [98].

$$\boldsymbol{sigmoid} \equiv \boldsymbol{f}(\boldsymbol{x}) = \frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{e}^{-\boldsymbol{x}}} \ldots\ldots\ldots\ldots\ 2.\ 13$$

  The output landscape of the sigmoid function is shown in Figure 2.5.

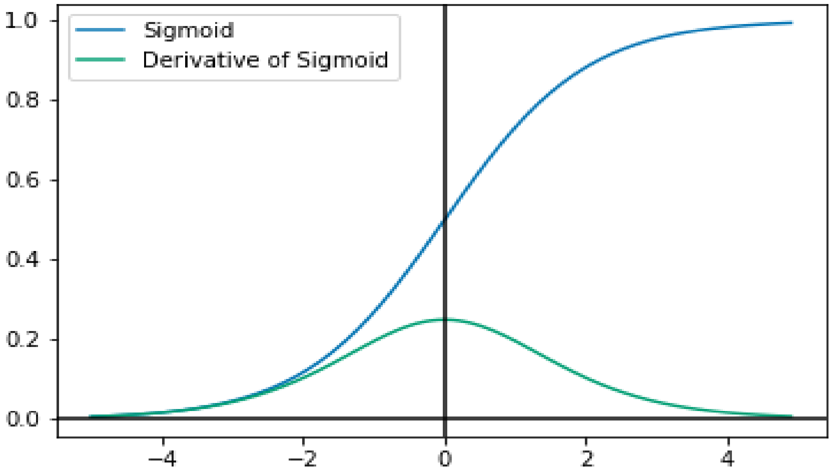

Figure 2. 5 Output Sigmoid Activation Function

The sigmoid functions have the advantage of being simple to grasp and are commonly utilized in shallow networks. Sharp damp gradients during backpropagation from deeper hidden layers to the input layers, gradient saturation, sluggish convergence, and non-zero-centered output are all important shortcomings of the Sigmoid AF, causing gradient updates to propagate in various directions. The Sigmoid and SoftMax AF differ primarily in that the Sigmoid is used for binary classification while the SoftMax is utilized for multivariate classification [98].

CHAPTER THREE

PROPOSED METHODOLOGY TO EXTRACT COHESIVE CANDIDATE SUMMARY

3.1 Introduction

This chapter presents the design of the proposed topic summarization system. This system can distinguish and extract salient important sentences of different sizes to form a cohesive candidate summary that fits with the golden summary. This chapter also provides three suggestions, the first one is a new clustering algorithm to collect unlabeled data, and the second creating a lexical chain based on sentences as a new idea different from the traditional lexical chain. This suggestion was achieved through the use of the proposed clustering algorithm in this dissertation too, and the third presents a new group combination of features extraction. All these suggestions are considered one of the important and necessary stages in the proposed system.

The proposed system comprises several stages that aim to extract a cohesive candidate summary starting from preprocessing techniques, embedding vector techniques, text clustering techniques, feature extraction, and sentences reduction techniques. In addition, this system has able of sorting candidate sentences based on time events. Therefore, this chapter aims to introduce innovative methods for the proposed system.

3.2 The Proposed System Layout

This section talks about the general layout of the proposed topic summarization system. The Proposed System combines two concepts clustering and classification can be coined (**Semi clustering unsupervised system based on extractive topic summarization**). It consists of several stages as stated in Figure 3.1 and each stage has role affective contributed in generate a cohesive candidate summary. Each stage of this system will be explained in the next sections.

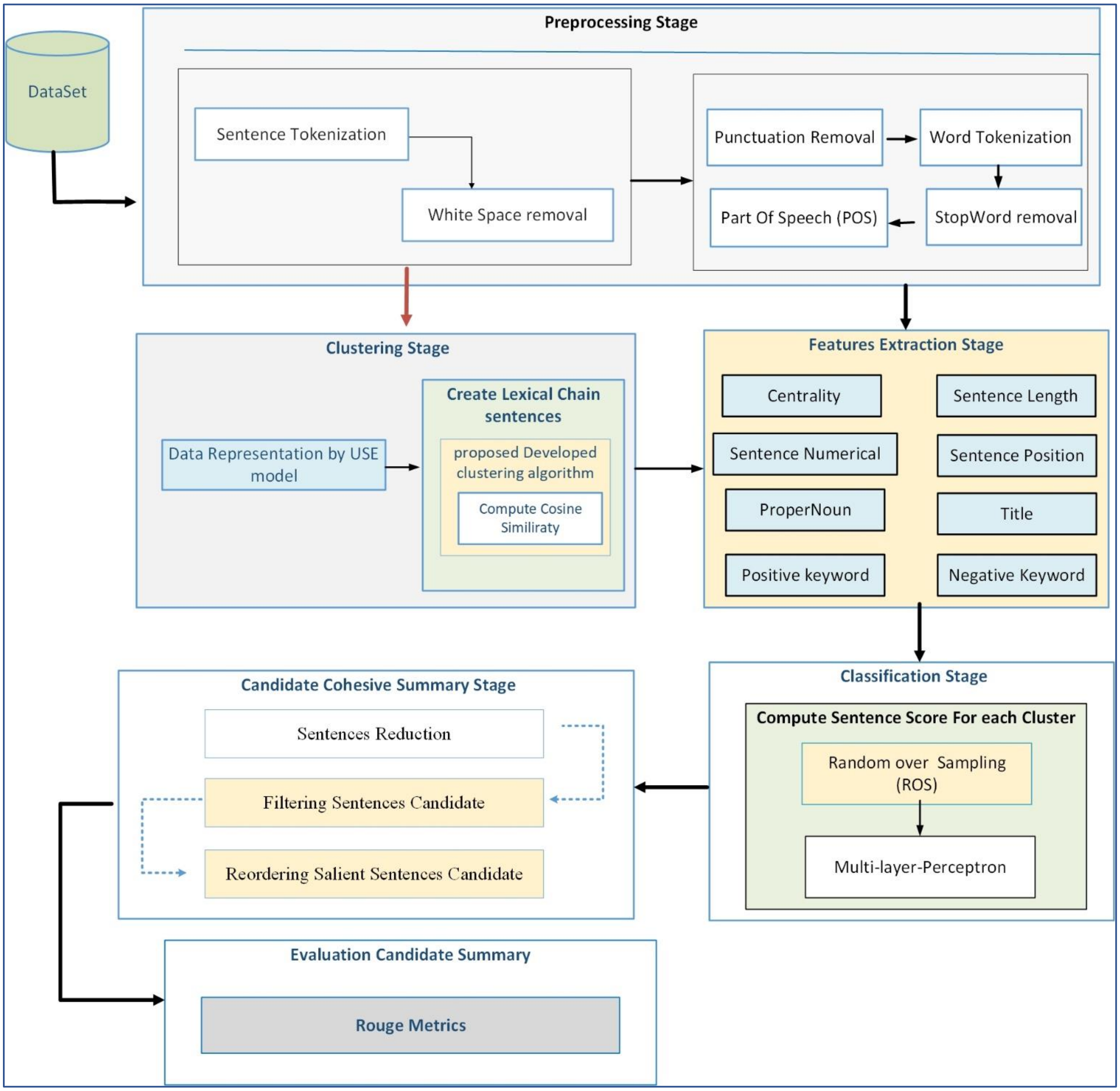
DataSet
Preprocessing Stage
Sentence Tokenization
White Space removal
Punctuation Removal
Word Tokenization
Part Of Speech (POS)
StopWord removal
Clustering Stage
Data Representation by USE model
Create Lexical Chain sentences
proposed Developed clustering algorithm
Compute Cosine Similiraty
Features Extraction Stage
Centrality
Sentence Length
Sentence Numerical
Sentence Position
ProperNoun
Title
Positive keyword
Negative Keyword
Candidate Cohesive Summary Stage
Sentences Reduction
Filtering Sentences Candidate
Reordering Salient Sentences Candidate
Classification Stage
Compute Sentence Score For each Cluster
Random over Sampling (ROS)
Multi-layer-Perceptron
Evaluation Candidate Summary
Rouge Metrics

Figure 3. 1 Summarization System Proposed Layout in General

3.3 Data Set

For example, in this dissertation used two datasets. The first named Document Understanding Conferences (DUC) 2002 special for text summarization, and it has 59 topics [99]. The DUC 2002 dataset is briefly described in Table 3.1. Each topic is comprised of a collection of documents (articles) $T_i = \{d_1, d_2, d_j, \ldots\ldots, d_n\}$, and each document d_j is comprised of a collection of sentences $d_j = \{S_1, S_2, S_k, \ldots\ldots, S_m\}$. Table 3.1 show describe for this dataset. Another dataset collected by picked news articles on certain topics from various news agencies over the internet, and extracted the golden summary for each topic using QuillBot software to ensure the efficiency of this system's performance. this dataset will be explained details in section 4.2.4.

In this work, all of these documents' sentences have been merged into a single file D^*. For the sake of simplification.

Table 3. 1 Description Dataset DUC 2002

Description	**DUC 2002 dataset**
Number of topics	59 (d061j through d120i)
Number of documents in each topic	~10
Total number of documents	567
Golden Summary length	200 and 400 words

3.4 Preprocessing Stage

Preprocessing is considered one of the main important steps in text processing in natural language preprocessing (NLP). This stage includes a set of procedures are: word tokenization, Punction removal, sentences tokenization, removal of stop words, part of speech (POS), and white space removal. Algorithm 3.1 explains the six procedures above and how applying them in this stage details. The main steps of this stage are described listed as follows:

1. Input multi documents for a certain topic taken from dataset and read their text.
2. At the beginning insert text completely and convert it into lower case words characters by using Lower Case Letters technique (LCLT) to avoid the variance between the words upper case and lower case that affects negatively on results preprocessing procedures later, also benefits for fast reading, and make all the sentences in one format.
3. Using Sentence Tokenization Technique (STT) to break text into sentences to benefited in clustering stage for encoded it by USE model which explained in section 2.7. later is segmented each sentence into words by using Word Tokenization Technique (WTT) for the purpose of completing the requirements calculation the features extraction stage utilized in this work.
4. Using the WhiteSpace Removal Technique (WSRT) to delete extra whitespace (double space) for each sentence by inserting all sentences produced by STT to make all sentences in a file have the same space between words. This step works to prepare the sentences in a manner appropriate to enter the clustering phase because the clustering phase depends on the first two steps, which end in this step. In addition, it is considered an important step for the rest of the preprocessing procedures that contribute to the extraction of features more accurately. These procedures will be explained in the next steps.
5. Deleting the punctuation existing in the sentences produced from $WSRT$ because it is insignificant information, by using the Punctuation Removal Technique (PRT) is beneficial in improving task performance.
6. The Word Tokenization Technique (WTT) is used to break down each sentence from PRT into words so that the rest of the preprocessing steps can meet the needs of the feature extraction stage.
7. The Stopword Removal technique (SWRT) is removing all words irrelevant produced from WTT such as “a,” “an,” and “the.” Less intuitively, commonly used

negation terms, such as “no” and “not,”, etc are also generally considered stop words.

8. The final step is to extract $part\ of\ speech\ (POS)$ for each word in a sentence depend on $SWRT$ results, like numeric words, noun words, Verb words, Adj words, etc. In this work benefit of POS is to identifying numeric words count in the sentence to be computed in the numeric feature, and noun words count in the sentence to be computed in the pronoun feature. These features will be explained in section (3.2.4).

Note that the cluster stage in the proposed system depends on the first three steps of the preprocessing techniques only. While the feature extraction stage made use of all preprocessing techniques to meet its requirements. In general, the fundamental background for the preprocessing procedures was discussed in detail in chapter two, in section (2.3).

```
Algorithm 3. 1: Preprocessing Stage.

Input: $Population\ Size\ N$ \\ N set of text taken from multi documents to the same topic.
Output: $SWRT$\\ $sentences\ without\ Stopword\ , POST$ \\ $part\ of\ speech$ for each word.
Begin
1. $LCLT \leftarrow \delta(\ lower\ case(N));$\\ $\delta$ represent function, convert capital into small latter.
    \\ The implementation of the procedure the $lower\ case\ \ function$
        return $N.lower()$ \\ final result which will save in $LCLT$
2. $STT \leftarrow \delta(Sentence\ Tokenization\ (LCLT));$ \\ $\delta$ represent function, $STT$ array of sentences.
    \\ The implementation of the procedure the $Sentence\ Tokenization\ \ function$
       return $sent\_tokenize(LCLT)$ \\ final result which will save in $STT$
3. $WSRT \leftarrow \delta(Extra\ space\ removal(STT\ ));$ \\ $WSRT$ saving the results of the function,
                                                  deleting extra whitespaces from each sentence.

4. $PRT \leftarrow \delta(\ punctuation\ removal(WSRT\ ));$\\ PRT saving the results of the function


5. $WTT_{ij} \leftarrow \delta(\ Word\ Tokenization(PRT\ ));$\\ $WTT$ array two-dimension saving results.
    \\ The implementation of the procedure the $Word\ Tokenization\ \ function$
    \\ Segment each sentence in $WST_i$ into words
    5.1 return $sent\_tokenize(LCLT)$ \\ final result which will save in $WTT$
```

```
6. $SWRT \leftarrow \delta\left(stop\ word\ removal(WTT_{ij})\right)$; \\ Delete the $stop\ word\ removal$
7. $POST_{ij} \leftarrow \delta\left(part\ of\ speech(SWRT_{ij})\right)$; \\ The implementation of the procedure the
                                                   $part\ of\ speech\ \ function$
8. End
```

3.5 Clustering Stage

This section discusses three significant aspects. The first consideration is how data is represented using the Universal Sentence Encoder (USE) model discussed in section 2.7. The second consideration is the scale utilized to compute the similarity of sentence vectors generated by the model. the third component, creates a lexical chain based on sentences coined (lexical chain sentence (LCS)) as a new contribution, which was utilized in this dissertation by proposing a novel clustering technique called (Developed clustering algorithm) and generating a graph for each cluster. Algorithm 3.2 explains the implementation mechanism steps to this stage in the proposed system, and these aspects (steps) will be discussed in further depth below.

```
Algorithm 3. 2 Clustering Stage ($WSRT$)

$\boldsymbol{Input}$: $WSRT$ \\ $WSRT$ array of the sentences without whitespace
$\boldsymbol{Output}$: $LCS$ \\ create Lexical Chains based Sentences
Begin
1. $ESV \leftarrow USE(WSRT)$; \\$ESV$ Embedding Sentence Vector
   \\ The implementation of the procedure is the universal encoder sentence(USE) model.
   \\ Convert each sentence into an embedding vector using the USE model.

        $\boldsymbol{For}\ i \leftarrow 0\ to\ size(\boldsymbol{WSRT})$

          $sent \leftarrow WST_i$

          $emmbeding\ array[i] \leftarrow embed(sent)$ \\ $embed$ Convert each sentence into an
        embedding sentence vector using USE model.
        $\boldsymbol{End}\ for$
      $\boldsymbol{return}\ emmbeding\ array$ \\ this final result will put in $ESV$
```

```
2. $SimiliratyMatrix \leftarrow CosinSimilarity\ distance(ESV\ )$; \\ see section 3.5.2
3. $LCS \leftarrow$ Developed $(SimiliratyMatrix)$;\\ see to algorithm 3.3 (Developed
   clustering)
4. End
```

3.5.1 Data Representation

This dissertation used the USE model as a tool to represent sentences as embedding vectors. The USE model is a technique working to convert text into a sentence embedding vector. It is characterized as representing sentences regardless of their length to a fixed length of size 512, it deals with the context of sentences and not a word because its method benefits in semantic matching so that takes the whole sentences instead of using wordnet or the thesaurus. Also, it solves a spars problem that exists in $\text{word 2 vector(w2v)}$ embedding vector technique because each sentence became has one dimension vector instead of many dimensions there more details are explained in section 2.7 about USE model.

3.5.2 Cosine Similarity scale

This scale is more popular used in text mining when asking for computing similarity among two sentences. The basic task of this scale is computing cosine distance semantic similarity between two sentences vectors generated by USE technique. Applying $equation\ (2.13)$ special with this scale which is mentioned in Chapter Two Section (2.8) to form matrix similarity $N \times N$ which will be useful to know the degree of similarity each sentence with the rest sentences, Table 3.2 shows this matrix produced from this process as an example. The Cosine Similarity scale in the normal case is neither interesting to semantic similarity nor the context. However, it cares about matching sentences degree with other sentences due to each vector being the Back Of Words (BOW). Therefore, the role of embedding concept and especially the USE model embedding sentences vector in this field appears because its concern to semantics and context when generating vector, thus making this scale intake consider semantic and similarity when computing as explained above.

Table 3. 2 Similarity Matrix Resulted from Computing Cosine Similarity for Sentences Vectors Generated By USE

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}	s_{11}	s_{12}
s_1	1	0.1	0.55	0.05	0.4	0.58	0.12	0.79	0.65	0.11	0.33	0.2
s_2	0.1	1	0.13	0.1	0.54	0.22	0.6	0.15	0.29	0.77	0.21	0.5
s_3	0.23	0.13	1	0.3	0.6	0.11	0.48	0.28	0.35	0.16	0.69	0.48
s_4	0.05	0.1	0.3	1	0.56	0.19	0.61	0.5	0.42	0.5	0.003	0.3
:	:	:	:	:	:	:	:	:	:	:	:	:
s_{11}	0.33	021	0.69	0.003	0.8	0.45	0.01	0.4	0.55	0.02	1	0.11
s_{12}	0.2	0.5	0.48	0.3	..	..	..	..	..	..	0.11	1

3.5.3 Create Lexical Chains Based on Sentence (LCS)

This work generates multi-chains based on sentences called Lexical Chain Sentences (LCS) or another expression it creates lexical cohesion based on the related sentences. This dissertation aims to establish a lexical chain sentence (LCS) depending on the closest semantic sentences or similar between sentences words that differ from the traditional Lexical Chain Word (LCW) which depends on the words. To create a lexicon proposed must find the clustering algorithm appropriate to help in achieving that. Therefore, the dissertation suggests a new clustering algorithm named (Developed clustering algorithm) that contributes in establish LCS proposed. Developed clustering is an algorithm that collects unlabeled data generally and constructs clusters occasion will explain in more detail in section (3.5.4).

In general, the LCS is a similar internal cluster principle so that each chain composes a set of sentences close to sense or repeated some of the share words among them. The idea of proposing LCS is to make the LCS deal with the problem of sentence sense disambiguation (SSD) and how to make it in the correct chain. Lexical sentence chains are created based on similar sentences sense and on another hand same based on the topic sentences of the document. The proposed LCS is constructed by computing the semantic distance among sentences by using the memetic between Universal Sentence Encoder (USE) method and Cosine Similarity distance coined USECS without using WordNet as LCW.

Usually, the LCS relationship that exists between sentences is extracted from $USECS$. Each sentence must belong to exactly one chain (cluster) when lexical chains are computed. An important advantage of the LCS is that it successfully overcomes the LCW challenge of word meaning demystification (WSD) by taking entire sentences without encoding sentences into words as occurs in traditional LCW. Each sentence has one sense, inverse to the words, which may have more than one sense. Thus, identifying the correct sentence sense will become easier. The LCS also prevents repeat sentences in more than one chain (cluster). Therefore, the LCS seeks to find the best method to collect sentences that will result in the longest and strongest chains. Figure 3.2 depicts the differences between the work LCW and the proposed LCS.

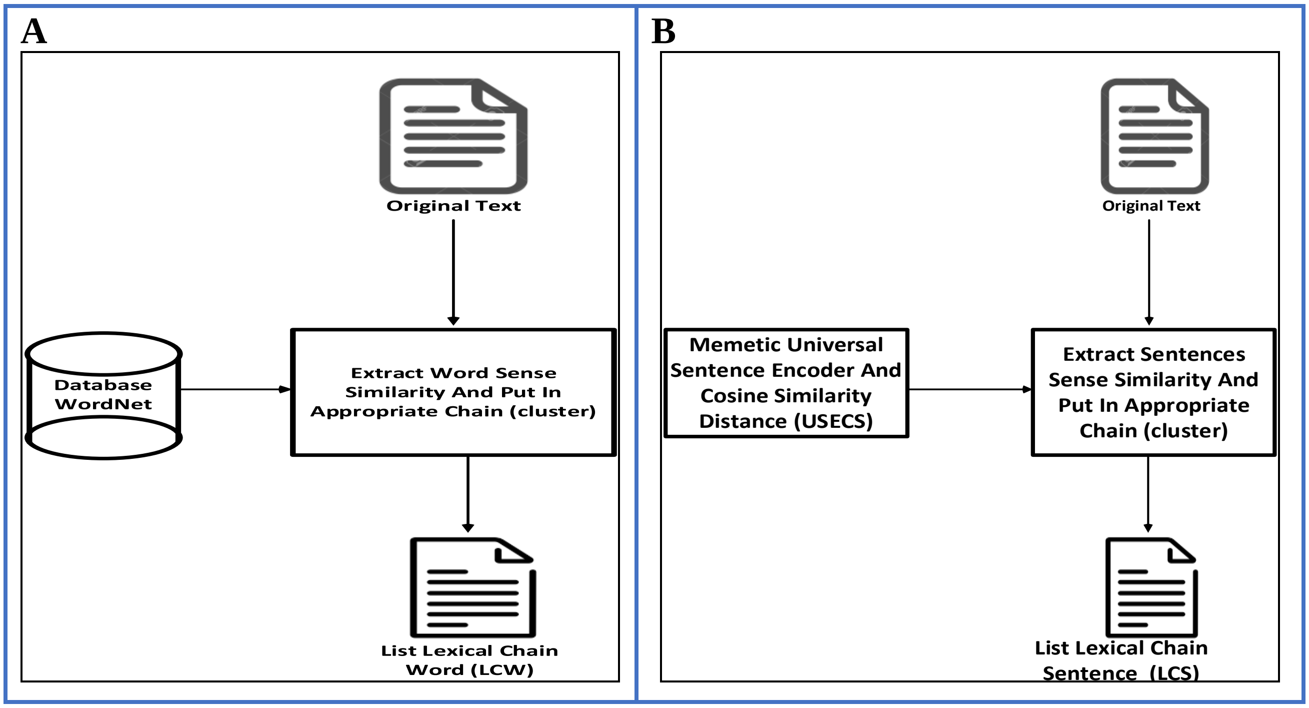
A
Original Text
Database WordNet
Extract Word Sense Similarity And Put In Appropriate Chain (cluster)
List Lexical Chain Word (LCW)
B
Original Text
Memetic Universal Sentence Encoder And Cosine Similarity Distance (USECS)
Extract Sentences Sense Similarity And Put In Appropriate Chain (cluster)
List Lexical Chain Sentence (LCS)

Figure 3. 2 A: Lexical Chain Word for Extracting Word Semantic Similarity from Text, and B: Lexical Sentences Graph for Extracting Sentences Semantic Similarity from The Text.

3.5.4 Proposed Developed Clustering Algorithm

A new algorithm named Developed clustering algorithm has been proposed in this dissertation. This algorithm plays an important role in grouping sentences according to sense or closet similar words for each cluster (Chain). The proposed algorithm has nine-phases as shown in Algorithm 3.3. In this section, the main steps of the proposed method are described listed as follows in detail:

1. The population is taken from a dataset.

2. Determine the threshold value previously and do not identify a number of the clusters because this algorithm is deciding the number of clusters optimality based on the threshold value automatically.

3. Using the USE model proposed by [44] for creating an embedding sentence vector for each sentence with a fixed length. this model is interesting with sentences (i.e.

context-based representation) only so that transforms each sentence completely into an embedding sentence vector.

4. Each sentence embedding vector (i.e., each sentence) will be put in a cluster to become a centroid (or center). Thus, will generation number of clusters equal the number of sentences.
5. Using Cosine similarity as stated in Equation 2.13 to compute the similarity degree between the center and the rest of the sentences. Thus, each center attracts sentences that are similar semantically through similar values resultant from the center and sentence which must be greater or equal to a threshold value and put in his cluster named $Cluster\ ESV_i$ where (ESV_i) is represent sentence vector index which becomes the center and Cluster ESV_i represent sub-matrix (cluster) including sentence numbers that attracted

6. Now, the proposed algorithm will reduce the number of clusters generated by considering each sentence attracted in the first cluster Cluster ESV_1 for example, cannot be selected or keep it as a center for another cluster because these sentences found a high affinity for the first cluster. Thus, each sentence existing in the Cluster ESV_1 must delete their cluster.

7. After completing the collection process of sentences in clusters and keeping them in a list named $Cluster\ ESV$, the proposed method must ensure that clusters of content are free of redundant the same sentence in more than one cluster. If such a situation exists, it compares similar values in the sentence in all clusters and the survival of this sentence in the cluster which has a higher similarity value than other clusters, then the similarity values of this sentence are deleted in the rest of the clusterable.

8. Checking the number of sentences in the cluster created. This method is required that the cluster content must be greater than two to avoid that being content in this cluster less than two after duplicate removal.

9. Now, there exist maybe sentences that have values of similarity with centers but are not compatible based on a threshold value, thus becoming these sentences called outliers. To include these outliers through taking similarity values each one of them with centers only and comparing among them and selecting a higher value and putting it in the cluster that belongs to that center. This list is considered certified lexical chain sentences. In this step, the process ends with creating the proposed LCS, or a set of clusters that cover all the sentences in the document. The characteristics Develop clustering algorithm characteristic are:

1. It is different from clustering algorithms in principle because it does not require identifying the number of clusters at the start, but it decides the number of clusters automatically based on the threshold value. Because the proposed algorithm allows to put a range of the threshold values. Then selecting optimal threshold based on the least positive evaluation of a series of evaluations, it was obtained through the DBI measure which is discussed in Chapter Two Section (2.11.1) for clusters generated by this algorithm. The algorithm 3.4 explains execution steps for choosing the best number of clusters. the results this process has been displayed in Section 4.2.1. While most cluster algorithms require identifying the number of clusters k in beginning like Kmeans algorithm.

2. It is similar to the hierarchical clustering algorithm in principle not requiring the number of k clusters at the beginning. Whereas it differs from the hierarchical clustering algorithm because when wanting to merge an item in a cluster, the proposed

method computes the distance between center and item only to decide merge or not without computing the distance between content cluster and item. The merge condition whether min\max depends on the data type used in work also according to the threshold value. Thus, this method is less expensive. while in hierarchical clustering merging process item with cluster occurs by computing scale between content cluster with an item then select min\max scale this depends on the data type and does not require to identify threshold value in order grouping points (sentences). Thus, hierarchical clustering considers more expensive.

3. In general, this method is suitable for clustering any numerical data type unlabeled.

4. It can abstract the properties of the Developed clustering algorithm the following:
   A. It depends on a distance measure like cosine distance.
   B. The number of clusters is based on the threshold value previous, thus it does not require an initial k centroid.
   C. It prevents the same item or sentences being repeated between clusters.
   D. Compute usage the Developed clustering algorithm time complexity is $O(n^2)$.

Therefore, this algorithm belongs to the environment of the clustering algorithms which depend on distance metrics. From the algorithms that classify in the same environment are Kmeans and hierarchical clustering algorithms for instance.

Algorithm 3. 3: Developed Clustering Method

```
$Input$: $Population\ Size\ WSRT$ \\ $WSRT$ array of the sentences without whitespace
$EmbeddingSentenceVector$ : $ESV$
$Threshold \leftarrow constant\ value$
$Output$: $TotalSentenceCluster$ \\ it is a matrix of two-dimension that contains set of sub-arrays
                                   inside, each array represents cluster includes the number of
                                   sentences.
1-Each $ESV$ is consider as center C in each cluster.
      $ESV = \{ ESV_1, ESV_2, \dots , ESV_n\}$
      $C = \{C_1, C_{2,\dots\dots,} C_n\}$ // list of size n where $C_i$ is empty set
      $size(ESV\ ) = size(C\ )$

2- The center selecting from ESV$_i$ this selection is sequential.
      $For\ i \leftarrow 0\ to\ size$(C)
           $For\ j \leftarrow 0\ to\ size$(ESV)
             $SimilartyResult \leftarrow Compute\ CosineSimilarty\ between\ C_i and\ ESV_j$;
             If ($SimilartyResult \geq Threshold$) then
                Cluster_matrix C$_i$ ← index (ESV$_j$);

             $Else$
                Next j
             $End\ if$
           $End\ for$
      $End\ for$
3- If center C is existence in other clusters, then remove it this center based on higher similarity with
   other.

Repeat step2 and step3 until being sentences which computable with the threshold at appropriate
clusters.
4- Filtering out duplicate sentences between any two clusters by keeping only the sentences with the
      highest similarity to its center.
5- Search for outlier sentences which incompatible with threshold and compare them with centers.
      inserting these sentences in one of $C$luster_matrix$_i$ which has the highest similarity with the
      center.
Step7: End;
```

Algorithm 3. 4: Method Selecting Best Number Of The Clusters

```
$Input: WSRT\ , T1, T2$ \\ $T1, and\ T2\ insert\ range\ of\ threshohs\ Let\ T1$ is a minimum value and
                              $T2$ is a maximum value.
$Output: ONC$ \\ **optimal number clusters**
    1. ESV ← $USE(WST\ )$;\\ **explained this step-in algorithm 3.2 (cluster stage)**
    2. $SimiliratyMatrix\ \leftarrow\ CosinSimilarity\ distance(ESV\ \ )$;
       $\boldsymbol{For}\ i\ \leftarrow 0\ to\ size\ (ESV)$
         $A \leftarrow ESV_i$ \\ first vector
          $\boldsymbol{For}\ j \leftarrow 0\ to\ size(ESV)$
              $B \leftarrow ESV_j$ \\ Second vector
              $numsum\ \leftarrow sum\ (dot\ productA\ with\ B)$
              $root1 \leftarrow square\ root(\ sumation\ A^2)$
              $root2 \leftarrow square\ root(\ sumation\ B^2)$
             $Z \leftarrow dot\ product\ root1\ and\ \ root2$
           $Similartyscore_{ij} \leftarrow division\ numsum\ by\ Z$   \\ $Similartyscore\ matrix$
       $\boldsymbol{retrun}\ similarityscore$
    3. $NC\ \leftarrow\ MB(SimiliratyMatrix)$; \\ explain in algorithm 3.3 (MB clustering)
    4. $\boldsymbol{If}\ (NC == 0\ )\ then$ \\ $NC$ number cluster
           $go\ to\ step\ 7$ ;
            $T1\ = T2$;
       $\boldsymbol{Else}$
           $go\ to\ step\ 5$
     $\boldsymbol{End}\ if$
    5. $DBI\ score\ \leftarrow\ DBI(NC)$ ; \\ DBI Davies Bouldin Index metric to evaluate $NC$
    6. $Matrix\ DBI[i] \leftarrow\ DBI\ score$ ; \\ Collecting DBI scores in the matrix DBI
    7. $\boldsymbol{If}\ (T1 == T2)$        $Then$
                $go\ to\ step\ 8$;
         $\boldsymbol{Else}$
           $T1 = T1 + 0.01$; \\ Increase $T1$ value to 0.01.
          $go\ to\ step\ 3$; \\ Repeat from step3 when the T1 value does not match with the $T2$ value.
         $\boldsymbol{End}\ if$
    8. $ONC\ \leftarrow\ Min(Matrix\ DBI)$;\\ $ONC$ optimal number of clusters by a select small value
    9. End
```

3.6 Feature Extraction Proposed

The benefit of this step is to identify the important sentences from the text. In this work, several features from the [23], [42], and [41] were selected and aggregated to form a new group to determine the significance of the sentence. This group is distinguished by combining the features of the informative score (important sentence) and the features of extracting sentences that are easy to understand named (readability features). The features informative scores are **Centrality, Positive Feature, Sentence Position Feature, Sentence Numerical Feature, and Negative Feature.** While, the features readability scores are **Proper noun, Sentence Length Feature, and Headline Feature.** These features have been described in detail in chapter two section (2.6). algorithm 3.5 describes implantation steps to extract the features for each sentence in the feature extraction stage.

Algorithm 3. 5: Mechanism Implementation Feature Extraction Stage

```
$Inputs : ESV\ , SWRT, WSRT\ , and\ POST\ results$

$Output: centrality, PN, SPF, SLF, HF, NF, PF, NF, and\ features\ score$

 1. $Centrality \leftarrow centrality\ feature(\boldsymbol{ESV});$
      \\ Compute the centrality feature for each sentence vector with the rest sentences
         vectors in the same document.
         m ← 0 \\ counter variable
         $\boldsymbol{For}\ i \leftarrow 0\ to\ size(\boldsymbol{ESV})$
             m ← $m + 1$
             $\boldsymbol{For}\ j \leftarrow 0\ to\ size(\boldsymbol{ESV})$
                 $\boldsymbol{If}\ ESV_i\ not\ equal\ ESV_j\ \ then$
                     $dist\ arry_{\ m} \leftarrow compute\ cosine\ distance\ between\ ESV_i\ and\ ESV_j$
                 $\boldsymbol{End}\ if$
             $\boldsymbol{End}\ for$
             $result\ sum \leftarrow summation\ of\ \ dist\ array_m$
         $centarlity\ array_{\ i} \leftarrow \ result\ sum$
         $\boldsymbol{End}\ for$
         $\boldsymbol{return}\ centarlity\ array$
 2. $\boldsymbol{PN} \leftarrow Proper\ nounfeature(\boldsymbol{POST});$
    \\ Compute Proper noun based on $POST$ results to identify the number of noun words for
       each sentence.
    $\mathbf{For}\ i \leftarrow 0\ to\ size(\mathbf{POST})$
      $list \leftarrow POST_i$
      $noun \leftarrow 0$ \\ is variable
      $\mathbf{For}\ j \leftarrow 0\ to\ size(list)$
        $word \leftarrow list_j$
        $\mathbf{if}\ word\ equal\ to\ unites\ \ then$ \\ unites nouns include ('NN', 'NNP', 'NNS', 'NNPS')
          $noun \leftarrow noun + 1$
        $\mathbf{End}\ if$
    $\mathbf{End}\ for$
    $POST\ array_i \leftarrow\ noun\ divsion\ by\ length\ list$
  $\mathbf{End}\ for$
  $\mathbf{return}\ POST\ array$
3.  $SPF \leftarrow Sentence\ Position\ Feature(SWRT\ \ );$
    \\ Compute Sentence Position Feature for each sentence for same document
      $\mathbf{For}\ i \leftarrow 0\ to\ size(SWRT)$
        $document\ \leftarrow\ SWRT_i$
        $total\_sentences\_at\_ecah\_doc \leftarrow len(document)$ \\ compute number of sentences.
        $\mathbf{For}\ j \leftarrow 0\ to\ size$(document)
          $indexsentence \leftarrow j + 1$\\ to avoid divide by zero.
          $indexorginal\ \leftarrow indexsentence - 1$
          $\mathbf{If}\ indexorginal\ \ equal\ to\ j\ \ \ \ then$
              $A \leftarrow\ 1\ /\ indexsentence)$
              $B \leftarrow\ (1\ /((total\_sent\_at\_ecah\_doc\ -\ indexsentence)\ +\ 1)))$
            $sent\_position\ =\ \max\ ((A, B)$\\ choose maximum value
          $\mathbf{End}\ if$
        $\mathbf{End}\ for$
      $\mathbf{End}\ for$
    $\mathbf{return}\ sent\_position$
4.  $SLF \leftarrow\ Sentence\ Length\ Feature\ (SWRT\ \ );$
    \\ Compute Sentence Length Feature for each sentence
    \\ 3.1 Find longer sentence in document
      $longer\ \leftarrow 0$
      $\mathbf{For}\ i \leftarrow 0\ to\ size(SWRT)$
        $sentence\ \leftarrow SWRT_i$
        $\mathbf{If}\ longer\ <\ len(sentence)\ \ then$ \\ len compute number words of sentence
          $longer\ sentence\ \leftarrow\ int(len(sentence))$\\ int is integer number
        $\mathbf{End}\ if$
    \\ 3.2 compute length score for each sentence
    $\mathbf{For}\ j\ \leftarrow 0\ to\ size\ (SWRT)$
      $document\ \leftarrow\ SWRT_j$
      $\mathbf{For}\ k\ \leftarrow 0\ to\ size(document)$
        $Length\ sentence\ \leftarrow\ len(document_k)$
```

```
          $length\ scor_{jk}\ \leftarrow\ length\ sentence/longer\ sentence$
        $\boldsymbol{End}\ for$
     $\boldsymbol{End}\ for$
    $\boldsymbol{return}\ length\ scor$
 5. $HF \leftarrow Headline(WSRT\ \ ,TEV)$;\\ title embedding vector (TEV)
    \\ Compute Headline Feature (HF) by finding the similarity degree between each
       sentence and headline in the document.
    $\boldsymbol{For}\ i \leftarrow 0\ to\ size(WSRT)$
      $document\ \leftarrow\ WSRT_i$
      $ESV\ document\ \leftarrow USE(WSRT_i)$
      $topic\ \leftarrow TEV_i$
      $\boldsymbol{For}\ k\ \leftarrow 0\ to\ size(ESV\ document)$
        $Eebedding\ sentence\ \leftarrow\ ESV\ document_i$
        $distance\ score\ _{ik} \leftarrow Cosine\ Similarity\ between\ Eebedding\ sentence\ and\ topic$
    $\boldsymbol{End}\ for$
   $\boldsymbol{End}\ for$
 $return\ distance\ score$
6. $NF \leftarrow\ Numerical\ feature\ (\boldsymbol{SWRT}\ \ );$
   \\ Counting the number of the numerical words for each sentence
   $\boldsymbol{For}\ i\ \leftarrow 0\ to\ size(\boldsymbol{SWRT})$
      $document\ \leftarrow SWRT_i$
      $\boldsymbol{For}\ j\ \leftarrow 0\ to\ size(document)$
        $count\ \leftarrow 0$
        $\boldsymbol{For}\ k\ \leftarrow 0\ to\ size(document_j)$
           $\boldsymbol{if}\ document_{jk}\ equal\ to\ units\ \ \ then$ \\ units like (1, one,2, two, sixteen, and etc.)
                $count\ \leftarrow count + 1$
          $\boldsymbol{End}\ if$
        $\boldsymbol{End}\ for$
        $Numerical\ score\ _{ij} \leftarrow count\ divided\ by\ length\ document_j$
     $\boldsymbol{End}\ for$
   $\boldsymbol{End}\ for$
   $\boldsymbol{retrun}\ Numerical\ score$
7. $PF \leftarrow\ Positive\ Feature(SWRT\ \ )$; \\ Compute Positive Feature (PF) to find the number
   of the positive words for the sentence.
 8. $NF \leftarrow\ Negative\ Feature\ (SWRT\ \ )$;\\ Compute Negative Feature (NF) to find the
   number of the negative words for the sentence.
 9. $features\ score\ \leftarrow collects\ results(centrality, PN, SPF, SLF, HF, NF, PF, and\ NF)$
```

```
   \\ collect all feature results and put them in the feature array.
   $\boldsymbol{v} \leftarrow \boldsymbol{-1}$
   $\boldsymbol{For}\ i \leftarrow 0\ to\ size\ (centrality)$
       $\boldsymbol{For}\ j \leftarrow 0\ to\ size(centrality\ _i)$
           $\boldsymbol{v} \leftarrow \boldsymbol{v+1}$
           $feature\ array_v \leftarrow (centrality_j, PN_j, SPF_j, SLF_j, HF_j, NF_j, PF_j, and\ NF_j)$
       $\boldsymbol{End}\ for$
   $\boldsymbol{End}\ for$
   $\boldsymbol{retrun}\ feature\ array$
10.    $\boldsymbol{End}$
```

3.7 Classification Stage (Finding Score for Sentences)

This stage is useful in finding the scores for sentences for distinguishing salient important sentences from each cluster and choosing it as candidate sentences to form a candidate summary. This stage starts after extracting the features for all sentences in all documents on the same topic. this dissertation used a Backpropagation Multilayer Perceptron Neural Network (BMPNN) to find a score for each Sentence. This stage has two sides. The first side is considered a complement to the clustering stage because it helps in evaluating sentences by giving scores for them in clusters and showing salient sentences for each cluster for choosing it. The second side is proving the power of the set of combination features used in this work, which helped in finding scores for the sentences.

To achieve the classification concept, this dissertation established a binary label of $('0'\ or\ '1')$ depending on the sentences of the golden summary for each topic existing in the dataset DUC 2002, so that every sentence matches the sentence of the Golden summary has a label $'1'$ and every sentence does not match the golden summary sentence has a label $'0'$ by computing cosine similarity between sentences golden summary and sentences file. Thus, became each sentence has a class label in the file to easiest apply the BMPNN method to find the sentence score.

Finally, this stage seeks to appear significant sentences in a summary that matches the sentences of the golden summary. To achieve this goal, the sentences that have only one label $'1'$ have been selected, and they represent the True positivity in the confusion matrix which depends on the predictions of the $BMPNN$. The $BMPNN$ is explained briefly in algorithm 3.6 by displaying all steps used in this stage. additional explained in section 3.7.2 in detail. Also, this stage used the $Random\ oversampling$ (ROS) method which to achieve the balance and obtain better results. This method is explained in section 3.7.1.

Algorithm 3. 6 Classification Stage

```
$Inputs: features\ score\ , and\ class\ label$
$Output\ : sentence\ score$
    1. $balance\ \leftarrow ROS(\ feature\ score, class\ label)$
     \\ Use the Random Over Sampling ($ROS$) method to imbalance the dataset for each cluster
     $label0\ \leftarrow 0$
     $label\ 1\ \leftarrow 0$
     $For\ i \leftarrow 0\ to\ size(class\ label)$
     $label\ 0\ \leftarrow label\ 0\ sum\ with\ number\ \ '0's\ in\ each\ cluster$
     $label\ 1\ \leftarrow label\ 1\ sum\ with\ \ number\ \ '1'$ s in $each\ cluster$
     $End\ for$
     $z\ \leftarrow label\ 0/label\ 1$
     $unbalance\ \leftarrow sum\ the\ label\ 0\ with\ the\ label\ 1\ then\ divided\ by\ z$
     $For\ j \leftarrow 0\ to\ size(class\ label\ \ )$
     $If\ class\ label\ equal\ to\ 1\ \ \ then$
     $list\ label\ 1\ \leftarrow feature\ score_j$ \\ list label 1 an array to put sentence has label 1
        $End\ if$
     $End\ for$
     $h\ \leftarrow\ -1$
     $h\ \leftarrow -1$
     $For\ k \leftarrow 0\ to\ size(feature\ score\ \ )$
        $h\ \leftarrow h + 1$
     $n\ \leftarrow n + 1$
        $If\ \ h\ equal\ to\ unblance\ \ \ \ then$
        $For\ i\ \leftarrow 0\ to\ size(\ list\ label\ 1)$
    $list\ ROS_n \leftarrow feature\ score\ _i$
    $class\ label\ ROS_i\ \leftarrow 1$
    $n\ \leftarrow n+1$
    $h \leftarrow -1$
    $\boldsymbol{End}\ for$
    $\boldsymbol{Else}$
    $list\ ROS_n \leftarrow\ feature\ score\ _k$
    $\boldsymbol{End}\ if$
    $\boldsymbol{End}\ for$
    $\boldsymbol{return}\ list\ ROS$
   2. $Normalization\ data\ \leftarrow standraize\ method\ (list\ ROS)$\\ to ensure unbiased data
   3. $Sentenc\ score\ \leftarrow BMPNN(\ \boldsymbol{Normalizatio\ data, class\ label})$
      \\ Find sentence score by building a backpropagation multilayer perceptron neural network.
     $\boldsymbol{For}\ i\ \leftarrow 0\ to\ size(\ \boldsymbol{Normalizatio\ data})$
        $input\ layer\ \leftarrow\ Normalizatio\ data_i$
      $hidden\ layer1\ \leftarrow Sigmoid\ (dot\ product\ between\ input\ layer\ and\ weight1)$
      $hidden\ layer\ 2\ \leftarrow Sigmiod\ (dot\ product\ between\ hidden\ layer1\ and\ weight2)$
      $output\ layer\ \leftarrow Sigmiod(dot\ product\ between\ hidden\ layer2\ and\ weight3)$
      $\boldsymbol{If}\ output\ layer\ not\ equall\ \boldsymbol{class\ label_i}\ \ then$
         Apply backpropagation process to modify result
      $\boldsymbol{End}\ if$
     $\boldsymbol{End}\ for$
```

3.7.1 Random Over Sampling (ROS) method

After became each sentence has a class label, observed the data used suffers from an imbalance in terms of the number of $'1's$ and $'0's$ in the target category classification, considering this is one of the challenges in the classification problems where the goal is to train a classifier that accurately predicts the class to which an observation belongs.

The sentences with a $'1's$ target class label are less than the number of $'0's$ target class label sentences. Therefore, to achieve balance and obtain better results, using the Random

Oversampling (ROS) method in training, this method explained detail in section 2.12. the ROS principle is achieved by proposing in this dissertation Equation 3.1 which is:

$$Z = \frac{Total\, no.of\ label\, '0'}{Total\, no.of\ label\, '1'}$$

$$\boldsymbol{S = \frac{Total\, no.of\ label\ '0' + \ Total\, no.of\ label\ '1'}{Z}} \ldots\ldots 3.1$$

Where Z is a result dividing the total number $'0's$ by the total number $'1's$ in the label column, and S represents a sum of the total number to $'1's$ with $'0's$ are divided by Z. Therefore, the sentences labeled $'1's$ are randomly repeated to ensure balance based on the S result.

3.7.2 Backpropagation Multi-Layer Perceptron Neural Network (BMPNN)

In this step, a score for each sentence is extracted using a backpropagation Multi-layer perceptron neural network. the features extraction are input to this network with bias, thus become input layer contain of nine neurons. This network consisting of three layers by applying the objective function in this model as equation (3.2) in this network. The first hidden layer ($L1$) includes seven neurons, the second hidden layer ($L2$) involves five neurons, and the last layer (L3) is output layer consists of one neuron is 0 or 1 of prediction as shown in Figure 3.3. while Figure 3.4 describes work at this step by displaying processes used in it event obtain sentences candidate summary. the architecture this network is learning rat =0.001, the Epoch=5000, and Batch size is 64. This method is good in finding the higher sentence score by multiplying the value of the features for each sentence with weights that are generated in this network and choosing them as sentences in the candidate abstract without going to extract the appropriate weights experimentally or manually as in the source [23] and [100]. The objective function proposed in this model is:

$$\boldsymbol{Score(S_i) = sigmoid\Big(\sum_{h=1}^{y}(sigmoid(\sum_{j=1}^{m} sigmoid(\sum_{i=1}^{n} F_i * W_i) * W_j\Big) * W_h)} \;.... \mathbf{3.2}$$

Where sigmoid represents the activation function used in the output layer, y represents the number of classes in the output layer $(\mathrm{L3})$, also sigmoid is an activation function used in two hidden layers, m represents the number of input hidden layer2 $(\mathrm{L2})$, n represents the number of input hidden layer1$(\mathrm{L1})$, $\mathrm{F_i}$ is number features as input data for each sentence, and W represents weights generated in each layer. Can abstracting steps find sentence score following:

Step1: Compute the Sigmoid activation function as Equation (2.20) which mentioned in section (2.6) in two hidden layers $(\mathrm{L1}, \mathrm{L2})$. The Equations from 3.3 into 3.8 display implement the Sigmoid activation function method between L1 and L2.

$$\boldsymbol{X1 = \sum_{i=1}^{n} F_i * W_i} \quad .... 3.3$$

$$\boldsymbol{L1 = sigmiod(X1)} \quad .... 3.4$$

$$\boldsymbol{X2 = \sum_{j=1}^{m} L1 * W_j} \quad .... 3.5$$

$$\boldsymbol{L2 = sigmiod(X2)} \quad .... 3.6$$

where X1 represents dot product between features and weights in hidden first layer L1, and X2 represents dot product between features and weights in hidden second layer L2.

Step2: in the output layer (L3) used the Sigmoid activation function as Equation (2.20) mentioned in section (2.6) to classify sentences according to class because this function can be used with binary classification.

$$\boldsymbol{X3 = \sum_{\mathrm{j=1}}^{\mathrm{m}} \mathrm{L2} * \mathrm{W_h}} \quad \mathbf{................ 3.7}$$

$$\mathbf{L3} \mathbf{=} \boldsymbol{sigmiod}(\mathbf{X3}) \quad \mathbf{................ 3.8}$$

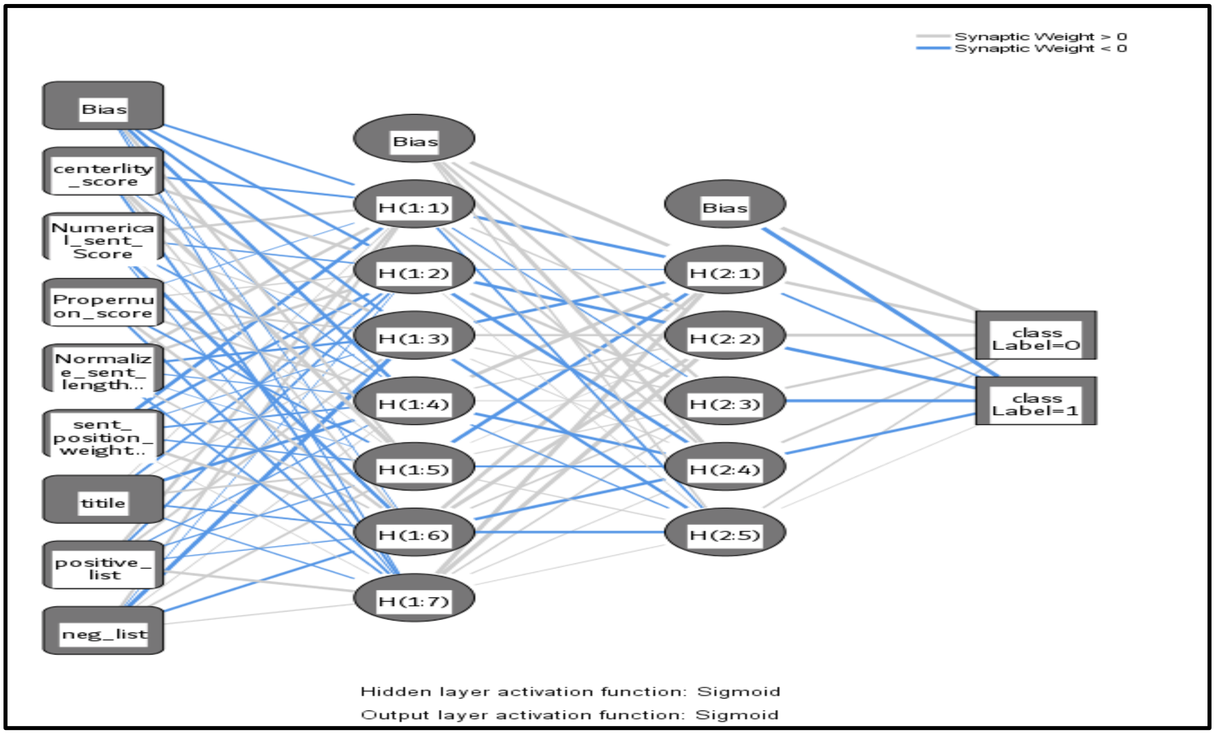

Figure 3. 3 Back Propagation Multi-Layer Perceptron Proposed for Giving a Score for Each Sentence

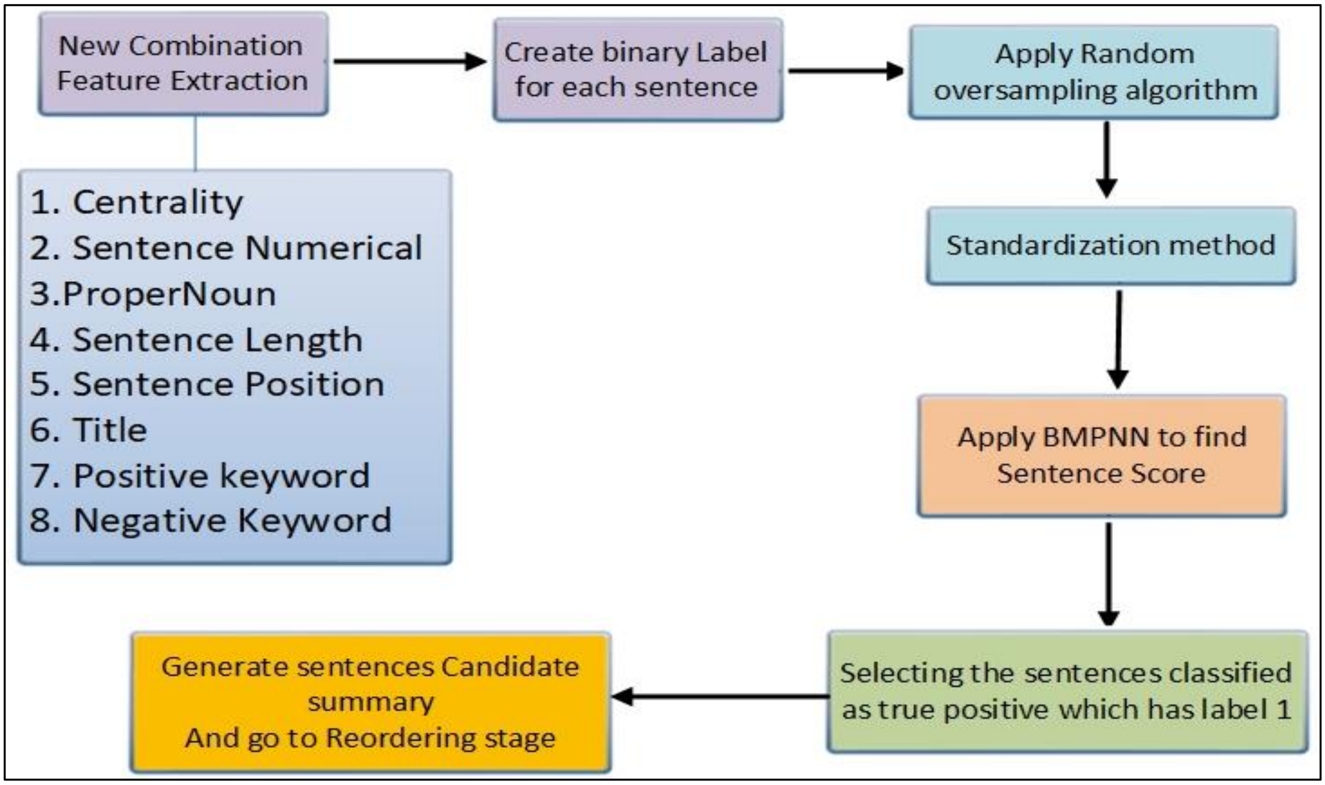
New Combination Feature Extraction
Create binary Label for each sentence
Apply Random oversampling algorithm
1. Centrality
2. Sentence Numerical
3.ProperNoun
4. Sentence Length
5. Sentence Position
6. Title
7. Positive keyword
8. Negative Keyword
Standardization method
Apply BMPNN to find Sentence Score
Selecting the sentences classified as true positive which has label 1
Generate sentences Candidate summary
And go to Reordering stage

Figure 3. 4 Layout BMPNN Processes

3.8 Candidate Cohesive Summary stage

To access into candidate cohesive summary, this stage considers three directions, the first one aims to extract a candidate summary from the sentences that match sentences of the golden summary as much as possible through choosing the sentences that have only label $'1'$, and they represent the True positivity in the confusion matrix which depends on the predictions of the BMPNN that are explained in the previous section; the second is the length candidate summary must be appropriate with a length of the golden summary by putting some of the conditions which next talking it in section (3.8.1), and the third is solving reordering problem of the sentences of the candidate summary which in section (3.8.2) will be explained. Algorithm 3.7 describes work at this stage as briefly.

Algorithm 3. 7: Candidate Cohesive Summary stage

```
$Input: sentence\ score, date\ and\ time, golden\ summary, SLF, and\ total\ sentence\ clustering$

$Output: candidate\ cohesive\ summary$

    1. $candidate\ summary \leftarrow Reduction(\ sentence\ score)$\\ reduction important sentences which
       classified true positive (TP) in confusion matrix
       \\ put all important sentences from all clusters in one cluster
       $n \leftarrow 0$
       $\textbf{For}\ i \leftarrow 0\ \ to\ size\ (sentence\ score)$
         $\textbf{For}\ j \leftarrow 0\ to\ size(sentence\ score_i)$
           $\textbf{If}\ \ sentence\ score_{ij}\ equal\ to\ class\ label_{ij}\quad then$
                   $repositary\ important\ sentence_n \leftarrow\ total\ sentence\ clustering\ _{ij}$
             $length\ score_n \leftarrow SLF_{ij}$ \\ sentence length feature score
             $n\ \leftarrow n + 1$
           $\textbf{End}\ if$
         $\textbf{End}\ for$
       $\textbf{End}\ for$
       $\textbf{return}\ repositary\ important\ sentence$

2. $posfilter\ \leftarrow$
   $lengthTotal(candidate\ summary, golden\ summary, all\ date, clock, centrailty, length)$

   \\ Compute the length of a candidate summary of words compared with the length of the golden
   summary.
```

```
    $length\ candidate\ \leftarrow len(candidate\ summary)$\\ compute number of words

    $length\ golden\ \leftarrow len(golden\ summary)$

    $\boldsymbol{If}\ length\ candidate\ larger\ than\ length\ golden\quad then$

        $\min length\ \leftarrow find\ minimum\ length\ feature\ score$

        $index\ \leftarrow index(length[\min length])$

        $\boldsymbol{Delete}\ length_{index}$

        $\boldsymbol{Delete}\ centrality_{index}$

        $\boldsymbol{Delete}\ date\ _{index}$

        $\boldsymbol{Delete}\ clock_{index}$

        $\boldsymbol{Delete}\ candidate\ summary_{index}$

    $\boldsymbol{End}\ if$

    Repeat this step until be candidate summary less or equal to golden summary

    $\boldsymbol{return}\ candidate\ summary$

3.  $candidate\ cohesive\ summary \leftarrow reordering(posfilter\ , all\ date, clock, centrality)$
  \\ Reordering the important sentences candidate resulting from step 2 to produce a cohesive summary.
  \\ 3.1 Before beginning the sorting process, save the date and location.
     $For\ i \leftarrow 0\ to\ size(all\ date)$
        $list\ pure\ date_i\ \leftarrow all\ date_i$
        $list\ pure\ index\ _i \leftarrow i$
     $End\ for$
  \\ 3.2 date sorting from the older to newest
     $Sort\ date\ \leftarrow sort(all\ date\ )$ \\ explained in section 3.8.2
  \\ 3.3 Saving new date location after sorting process
     $n\ \leftarrow 0$
     $For\ i\ \leftarrow 0\ to\ size\ (sort\ date)$
       $For\ j \leftarrow 0\ to\ size(list\ pure\ date)$
         $If\ sort\ date_i\ equal\ to\ list\ pure\ date_j\quad then$
           $list\ index_n\ \leftarrow list\ pure\ date_j$
           $n \leftarrow n + 1$
         $End\ if$
       $End\ for$
     $End\ for$
```

```
\\ 3.4 Check dates when exist frequent same date
    $f \leftarrow 0$
    For $i \leftarrow 0$ to size(all date)
      date ← all $date_i$
      If date Exist in list avoid_overloop then
          Next i
      Else
          $z \leftarrow 0$
          $k \leftarrow 0$
          For $j \leftarrow 0$ to size(all date)
            If all $date_i$ equal to all $date_j$
                $z \leftarrow z + 1$
                list index similarity $date_k$ ← list $index_j$
                $k \leftarrow k + 1$
             End if
          End for
          If z lager or equal than 2 then
            sim ← 0
            list $avoid\_overloop_f$ ← date
            $f \leftarrow f + 1$
            For $m \leftarrow 0$ to size(list index similarity date)
                \\ check when exist two or more date similarity have different clock.
                If clock candidate $list_m$ not exist in list frequent clock then
                    clock candidate $list_m$ append into list frequent clock
                Else
                    \\ compute number of frequent to that clock then adding to list frequent clock
                      $sim \leftarrow sim + 1$
                      clock candidate $list_m$ append into list frequent clock
                End if
            End for
            If sim equal to len(list index similarity date) then
                Clear contains the clock candidate list
                For $d \leftarrow 0$ to size (list index similarity date)
                      $centrality_d$ Append into list centrialy temparry
                End for
                list centrialy after sort ← Sort (list centrialy temparry) \\ sort Ascending
                For $L \leftarrow 0$ to size(list centrialy after sort)
                        date final ← list pure date [put index for $centraity_L$]
                        final sentences ← filter candidate[put index for $centraity_L$]
                        final clock ← clock[put index for $centraity_L$]
                End for
            Else
              list after sort ← Sort (list centrialy temparry) \\ sort Descending
              For $L \leftarrow 0$ to size(list after sort)
                 w ← index clock $_L$
                $date\ final \leftarrow list\ pure\ date_w$
                $final\ sentences \leftarrow filter\ candidate_w$
                $final\ clock \leftarrow clock_w$
              $End\ for$
          $End\ if$
        $return\ final\ sentences$
4. End
```

3.8.1 Filtering candidate sentences

The length of the candidate summary must be suiTable with the length of the golden summary for implementing the evaluation process correctly later. however, the length of the candidate summary maybe not be compatible with the length of the golden summary existing in the dataset used. Due to existence candidate sentences one or more effects on increased length summary produced. Thus, must put condition-based on one of the features extractions used in this dissertation to overcome this problem is the Sentence Length feature. This condition is removed sentence based on the smallest sentence length feature value Then the length of the candidate sentences summary is recalculated candidate summary. In the case candidate summary if still is the largest from the golden summary, then repeated and apply this condition above again. The end of the goal of the filtering process is the issuance version of the candidate summary is convenient with a golden summary in terms of length.

3.8.2 Reordering Sentences of Candidate Summary

The reordering procedure for produced summary is considered one of the problems relevant to text summarization that must be solved. although conducting many experiments in choosing the feature based on which the sentences are arranged, in the end, the centrality feature was relied on to rearrange the sentences. However, to achieve more consistency and cohesive, this dissertation success in solving this problem by highlighting the arrangement (reordering) of the candidate sentences according to the date and time.

The date and time can consider an additional feature extracted from each document then put the front of each candidate sentences date and time that belong to it. This stage takes into consideration similar dates for candidate sentences through establishing two checkpoints. The first checkpoint employed to ensure whether if similar dates or not. If this checkpoint finds a similar date, then it groups these sentences that have similar dates after those rearrangements these sentences based on time. The second checkpoint works by reordering candidate sentences that have the same date and time based on the centrality feature descending so that the sentence that has max centrality value is being before other, after complete this process is added into reminding sentences candidate summary.

Thus, we will obtain on the candidate summary is characterized by consistency (cohesive) and synchronized events in the subtraction. this dissertation is interested in date and time because it deals with articles news, thus the date and time forming factor is important in a rearrangement of the sentence's locations. Figure 3.9 describes the operations performed at this stage as a flowchart. Here end generating process cohesive candidate summary after passing the several stages, therefore must evaluate it with a golden summary in the next stage using Rouge metrics. Note the evaluation stage will discuss in detail in chapter four.

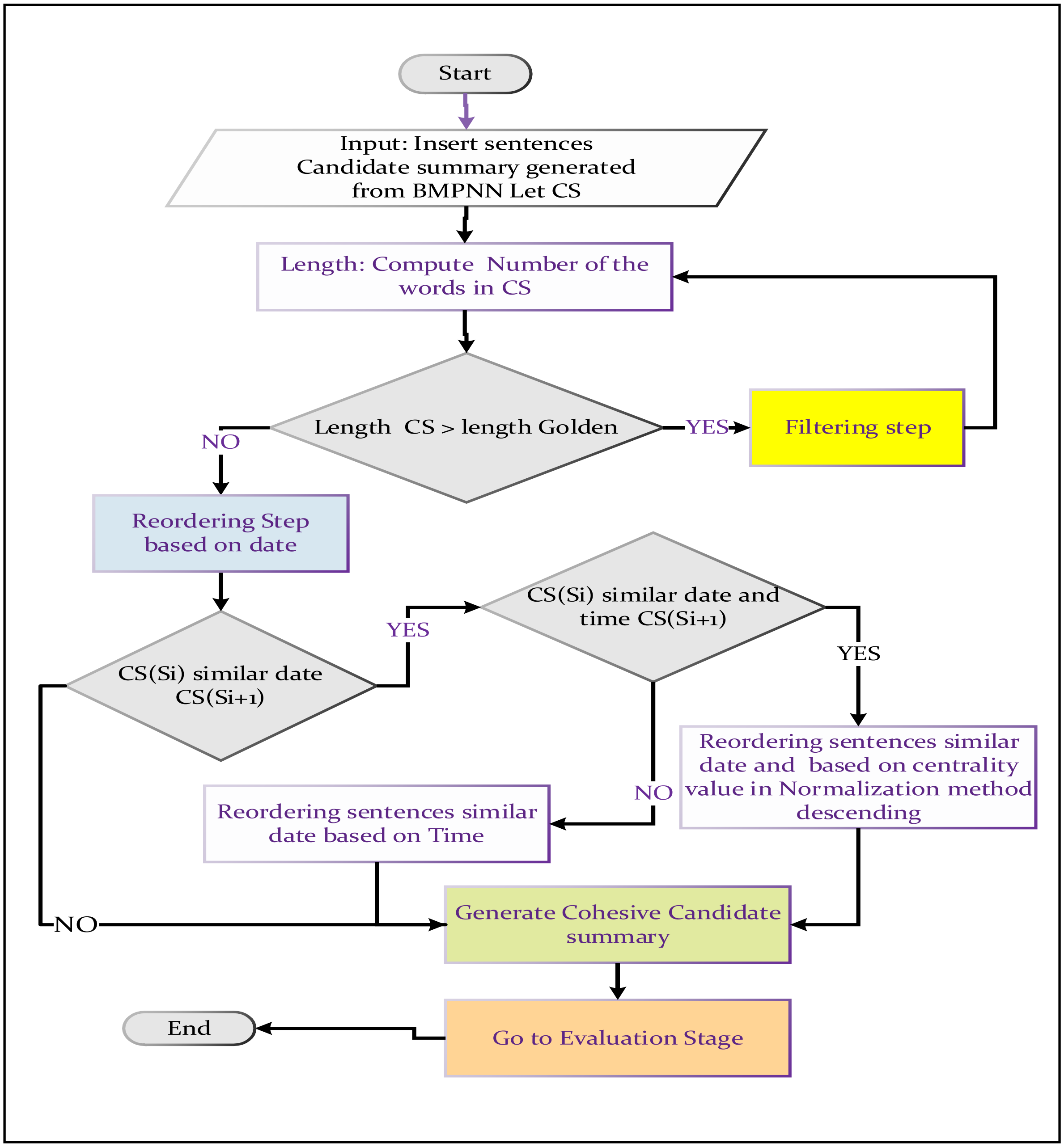
Start
Input: Insert sentences Candidate summary generated from BMPNN Let CS
Length: Compute Number of the words in CS
Length CS > length Golden
YES
Filtering step
NO
Reordering Step based on date
CS(Si) similar date CS(Si+1)
YES
CS(Si) similar date and time CS(Si+1)
YES
Reordering sentences similar date and based on centrality value in Normalization method descending
NO
Reordering sentences similar date based on Time
NO
Generate Cohesive Candidate summary
Go to Evaluation Stage
End

Figure 3. 5 Reordering Step

CHAPTER FOUR

EXPERIMENTAL RESULTS AND DISCUSSIONS

4.1 Introduction

This chapter compares the performance of the proposed MB clustering algorithm to the well-known techniques, the Kmeans algorithm, in terms of the number of clusters created. In addition to evaluating the correlated relationships between the number of clusters generated by K-means and the developed clustering algorithm only using the Davies Bouldin Index (DBI), this chapter also displays the evaluation results of the cohesive candidate summary produced by the topic summarization system proposed by Rouge measures.

4.2 Experimental Results

The experiments were conducted under Windows 10 Pro operating system, Intel(R) Core (TM) i7-6920HQ CPU @ 2.90GHz 2.90 GHz, 16 GB random access memory, and 64-bit system type. This section presents the performance evaluation of the MB algorithm proposed and the summarization system proposed. The evaluation is presented in terms of $F\_score, Precision, R, DBI\_score,$ and ACC. The language programming used in this dissertation is Python.

4.2.1 Developed Clustering Algorithm Performance Evaluation

This section presents the results of the Developed Clustering Algorithm proposed compared with the K-means in terms of performance and the number of clusters generated through a series of threshold values that were imposed as a period starting from 0.5 to 1.0 as showed in Table 4.1. The reason for starting from a value of 0.5 is due to many of the experiments conducted in this work until found that the similarity of the sentence begins with convergence from this value and upwards. This was proven when calculating the similarity between the outliers points and the centers of the clusters. Therefore, when it is less than half the similarity is almost non-existent or very weak. While the reason for determining a value of 1.0 as the end of the period is mean that the sentences that reach similar to this value are means high match state (i.e., may be repeated in another document). Thus, can benefit from this case to knowledge of redundant sentences by removing them and keeping only one of them.

Due to the large volume of data in the number of topics in the DUC 2002 dataset, which reaches 59 topics as described in detail in Section 3.2.1. Also, the huge volume of results, six of the topics were selected randomly. the topics encoded are 064,078,087,090,105, and 119 to present their results. Note, each encoded indicates behind the title for news.

Tabel4. 1 Number Of Clusters Generated From proposed Developed clustering algorithm

TOPIC	Developed Clustering Algorithm	
	NUMBER OF CLUSTERS	THRESHOLD
64	14	0.5
	10	0.51
	11	0.52
	10	0.53
	7	0.54
	5	0.55
	4	0.56
	2	0.57
	2	0.58
	2	0.59
	16	0.63
	13	0.64
	9	0.65
	8	0.66
	8	0.67
	7	0.68
	7	0.69

	5	0.7
	3	0.71
	2	0.72
	2	0.73
	2	0.74
	2	0.75
78	23	0.5
	22	0.51
	19	0.52
	19	0.53
	16	0.54
	14	0.55
	10	0.56
	9	0.57
	7	0.58
	6	0.59
	6	0.6
	6	0.61
	5	0.62
	3	0.63
	3	0.64
	2	0.65
	2	0.66
87	29	0.5
	25	0.51
	21	0.52
	21	0.53
	21	0.54
	18	0.55
	14	0.56
	14	0.57
	12	0.58
	11	0.59
	6	0.6
	5	0.61
	5	0.62
	4	0.63
	4	0.64
	3	0.65
	3	0.66
	2	0.67

90	20	0.5
	19	0.51
	14	0.52
	13	0.53
	14	0.54
	9	0.55
	12	0.56
	9	0.57
	6	0.58
	9	0.59
	11	0.6
	10	0.61
	7	0.62
	8	0.63
	5	0.64
	5	0.65
	5	0.66
	5	0.67
	4	0.68
	3	0.69
	2	0.7
	2	0.71
	2	0.72
	2	0.73
	2	0.74
105	15	0.5
	14	0.51
	12	0.52
	11	0.53
	10	0.54
	8	0.55
	6	0.56
	5	0.57
	5	0.58
	4	0.59
	2	0.6
	2	0.61
	2	0.62
119	13	0.5
	14	0.51
	12	0.52

	12	0.53
	10	0.54
	9	0.55
	7	0.56
	6	0.57
	5	0.58
	4	0.59
	3	0.6
	2	0.61
	2	0.62

Above Table 4.1 explains that the number of clusters in the proposed method in all topics. But the number of the clusters may be frequenting and this does not mean frequent same sentences or same centers. Due to it based on condition, mean a set of sentences compatible with one center. Thus, become this a set within a content this center. Therefore, the number of clusters is different as when $T = 0.52$ the number of the clusters is 19 whereas when being $T = 0.6$ the number of clusters is 6 for example in topic 078. Also, the series of threshold values assumed within the period may be un reach the period end which is 1.0 during generating a number of the clusters such as the topic 064 last threshold value is 0.75, topic 078 the last threshold value is 0.66, topic 087 the last threshold value is 0.67, topic 090 the last threshold value is 0.74, topic 105 the last threshold value is 0.62, and topic 119 the last threshold value is 0.62 because each topic independent from other in terms content and also the proposed method does not create clusters because the sentences maybe are not compatible with the threshold value specially or with the condition generally.

When compare the proposed algorithm with the $\mathbf{k-means}$ algorithm as shown in Table 4.2 by taking the same number of clusters generated in the two algorithms and subjecting them to $\mathbf{the\ Davies\ Bouldin\ Index\ (DBI)}$ scale. The $\mathbf{DBI}$ measure is utilized to evaluate cluster content and cluster relationships among them or the other word, it is used on two sides the first to assess clusters and knowledge of the strong relationship between them, and another side is to evaluate correlation content between them for each cluster. In the end, it gives a score. This score when to be a positive and low value it is considered good and indicates that this method is strong and better, and more detail about this measure was explained in section (2.11.1). The goal of using this measure is to evaluate the power of these algorithms in internal clustering. Note, this work has taken all topics from the DUC 2002 dataset to display how distributing sentences in the clusters is based on threshold value in the proposed method or based on the number of k as in the $\mathbf{k-means}$ algorithm. Figure 4.1 shows a brief description of the above six topics to increase understanding and convergence of viewpoints through comparison results.

Tabel4. 2 Evaluation K-means and proposed method by using DBI measure.

Topic	Number Of Cluster	DBI_Score for Developed clustering	DBI _Score for K-means
064	14	0.685265	2.789
	10	0.631362	2.948
	11	0.535524	2.860
	7	1.146702	3.290
	5	2.816969	3.117
	4	4.618549	3.309
	2	28.05815	3.820
	16	0.748338	2.807
	13	1.451087	2.914
	9	2.870145	3.310
	8	1.803425	3.368
	3	17.64205	3.522
78	23	0.167069	2.564
	22	0.197423	2.513
	19	0.294197	2.626
	16	0.250165	2.732
	14	0.395143	2.959
	10	0.768576	3.093
	9	0.839562	3.207
	7	1.895056	3.259
	6	3.023633	3.447
	5	4.655018	3.508
	3	20.08902	3.520
	2	41.89374	3.872

87	29	0.116223	2.354
	25	0.203342	2.427
	21	0.350926	2.565
	18	0.19916	2.825
	14	0.353845	3.008
	12	0.459526	3.133
	11	0.532208	3.131
	6	2.89072	3.376
	5	3.692883	3.499
	4	6.559097	3.533
	3	12.01871	3.679
	2	34.2646	3.892
90	20	0.307027	2.599
	19	0.331096	2.644
	14	0.682036	2.822
	13	0.809205	2.700
	9	1.326829	2.980
	12	0.575502	2.779
	6	2.568761	3.268
	11	0.568075	2.928
	10	0.670942	2.871
	7	1.539255	3.073
	8	1.173263	3.087
	5	3.368386	3.270
	4	5.112887	3.513
	3	10.62775	3.508
	2	28.57484	3.416
105	15	0.476057	2.872
	14	1.059623	2.806
	12	0.500509	2.787
	11	0.646664	2.951
	10	0.688129	3.163
	8	1.297875	3.249
	6	2.352609	3.300
	5	3.336893	3.250
	2	39.72709	3.748
119	13	0.790258	2.654
	14	0.644743	2.635
	12	0.985493	2.775
	10	1.539851	2.848
	9	1.982295	3.038

	7	3.403959	3.063
	6	5.62241	3.362
	5	7.543996	3.456
	4	11.89822	3.505
	3	35.13887	3.759
	2	27.33936	4.046

Table 4.2 displays the evaluation of the results $\mathrm{k}-\mathrm{means}$ algorithm and the proposed method using DBI measure. Most experiments in evaluating the number of clusters generated totally in the proposed method are successful in grouping clusters with more correlation and efficiency than the k-means algorithm. The DBI scores for the algorithms when being an algorithm lower score than another algorithm which means the algorithm is good. Since the DBI score when being low this means a good indication. This Table is contained six topics coined $064{,}078{,}087{,}090{,}105, and\ 119$. Each topic contains evaluations between the proposed algorithm and $\mathrm{K}-\mathrm{means}$ algorithm by using the DBI measure.

The results of topic 064 show that the proposed method is not successful with $k-mean$ when the cluster number is $2{,}3, and\ 4$, but in the remaining clusters it was successful. Topic 078 shows the proposed method better than the $\mathrm{k}-\mathrm{means}$ method in all numbers of clusters except cluster numbers $2{,}3, and\ 5$. In topic 087 the proposed method success only in cluster numbers $29{,}25{,}21{,}18{,}14{,}12{,}11, and\ 6$ while in the remaining clusters to the same topic is failed like $2{,}3{,}4, and\ 5$. In topic 090 the proposed method is advancing on the k-mean algorithm except for cluster numbers $2{,}3{,}4, and\ 5$. In topic 105 the proposed method outperforms the $K-means$ algorithm in all clusters except cluster numbers $2, and\ 5$. Finally, in topic 119 the proposed method in most cases is outperformed on the $\mathrm{K}-\mathrm{means}$ algorithm except for some of the clusters which are cluster numbers $7{,}6{,}5{,}4, \mathrm{and}\ 2$ did not have good luck in those clusters.

According to the DBI scores shown in the Table above, when the number of clusters reaches five or fewer, the DBI score examined for Developed clustering is a high positive score. This means that Developed performance in the correlation content and relationships between clusters begins to deteriorate, as illustrated in Figure 4. 1, due to the nature of the dataset utilized, which is text data that looks for similarity using the cosine similarity scale. The dissertation advises employing the dataset of interest in another field and using a distance metric such as Euclidean, for example, to see if the performance improves or vice versa, which can be recognized in future investigations.

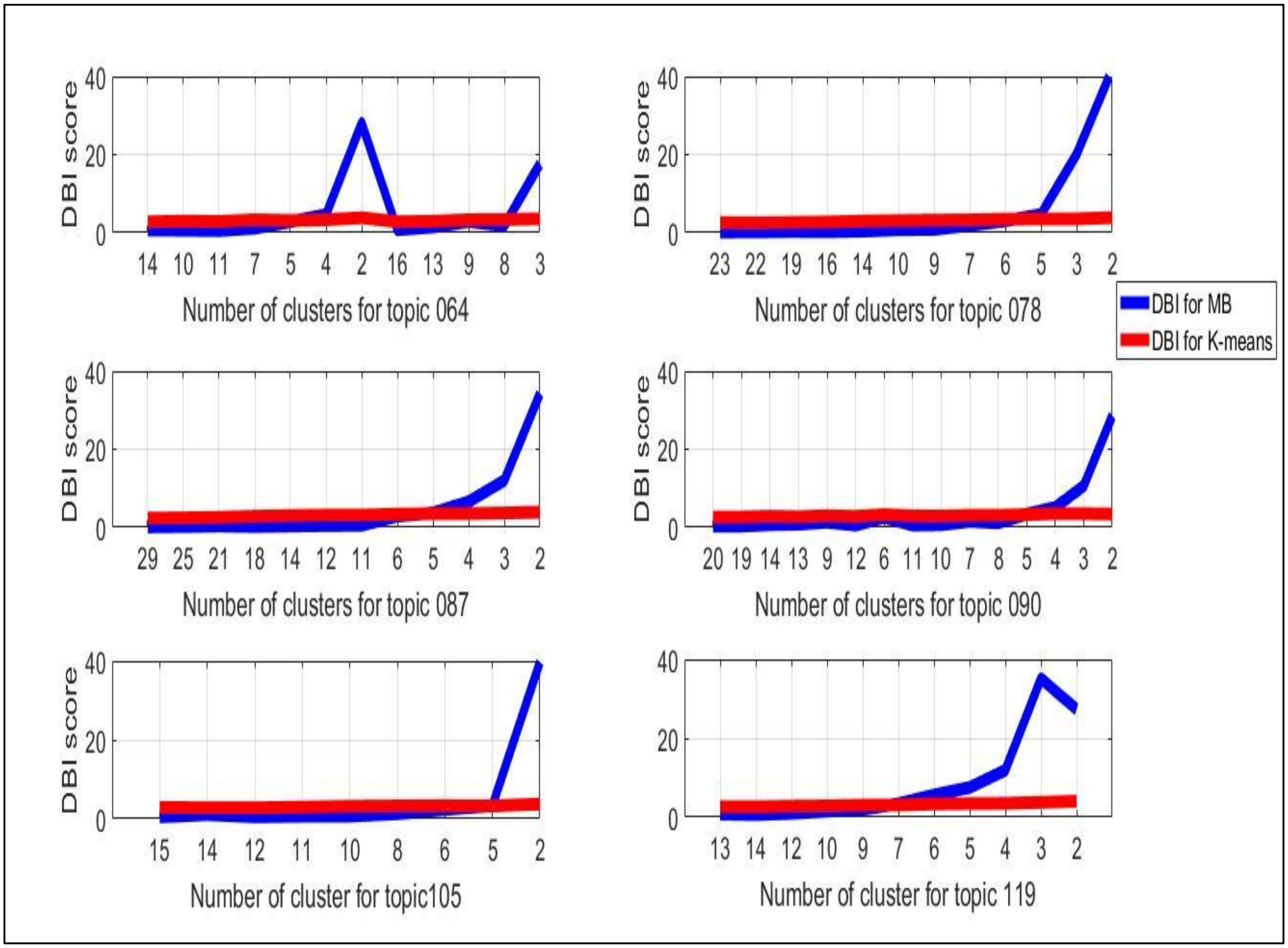

Figure 4. 1 Evaluations DBI Measure Between Developed clustering And K-Means Algorithm

This scale was also used to choose the best number of clusters, represented by the threshold value generated for each topic the DBI scale chooses the best group of the clusters resulting from a certain threshold value based on the lowest positive evaluation score because it considers the number of the clusters are good when achieves the lowest positive score during evaluation. Thus, can obtain the threshold value as an optimal value depending on the degree of its evaluation to the group of the clusters which generated it, as shown in Table 4.3, and its adoption in the construction of Lexical Chain Sentences (LCS) proposed.

Tabel4. 3 Optimal Threshold Values Selected By The DBI Method

Topic	Optimal Threshold	Number of Clusters
064	0.52	11
078	0.5	23
087	0.5	29
090	0.5	20
105	0.5	15
119	0.51	13

The above Table presents the best optimal threshold values that were selected by the DBI scale by evaluating the groups of clusters generated from these values and which depending on it in creating LCS proposed and access into forming a candidate summary later. The various in choose optimal thresholds in all topics back to less score evaluation achieved during calculation by DBI measure.

It is always better if the small threshold is not considered. For instance, the optimal threshold in topic 064 is 0.52, which means the number of clusters in it is eleven, while when the threshold is 0.5, it has fourteen clusters, which is not considered optimal because this measure may reveal strong links and correlation among clusters when there are only 11 clusters.

Therefore, it can be concluded that the evaluation metric showed the proposed algorithm succeeded in evaluation impressively in many experiments when the number of groups is small, regardless of the presence of some minor failures. Thus, it can be said that each algorithm has successes and failures. The proposed Developed clustering algorithm is considered the best compared with the $\mathrm{K-means}$ algorithm in terms of relationships and correlations between clusters.

4.2.2 Extract Salient Important Candidate Sentences

The experiments in the BMPNN focused on the aspect of classification and the process of prediction in distinguishing the important sentences corresponding to the golden summary before and after using the ROS method. The reason for using this method is to create a balance between data points exactly among class labels as mentioned in section 3.2.5.1. Observing that the $\mathrm{DUC}\ 2002$ dataset has a very small number of class labels $‘1’s$ compared with $‘0’s$ this means there exists an unbalance as shown in Table 4.4. Also, the ROS played an important role in the training process, and it was reflected positively in the testing process of the neural network, as shown in Tables 4.5 and 4.6. On the other hand, this section focuses on the process of selecting the important sentences that carry the label ‘1’.

Tabel4. 4 The Number of The Class Labels between '1' and Label '0' Before and after Using the ROS Method

Topic	Summary of words	Method	Total Sample	No. of class label “0”	No. of class label “1”
064	200	Before use ROS	182	174	8
		After use ROS	342	174	168
	400	Before use ROS	182	168	14
		After use ROS	336	168	168
078	200	Before use ROS	262	257	5
		After use ROS	477	257	220

	400	Before use ROS	262	246	16
		After use ROS	446	246	240
087	200	Before use ROS	264	257	7
		After use ROS	495	257	238
	400	Before use ROS	264	250	14
		After use ROS	488	250	238
090	200	Before use ROS	233	229	4
		After use ROS	417	229	188
	400	Before use ROS	233	223	10
		After use ROS	443	223	220
105	200	Before use ROS	235	228	7
		After use ROS	438	228	210
	400	Before use ROS	235	225	10
		After use ROS	425	225	220
119	200	Before use ROS	187	182	5
		After use ROS	342	182	160
	400	Before use ROS	187	178	9
		After use ROS	349	178	171

Table 4.4 shows the wide gap between the sentences scored as the class label '1' and sentences rated as the class label "0" this was before the use of the ROS method. This Reflects negatively because the network was not able to train appropriately and accurately. Thus, the network favors sentences with the class label "0" since their number is greater than that of sentences with the class label $'1'$. For example, in $Topic\ 064$, the important sentences that carry a label $'1'$ to cover a summary of 200 words are 8 sentences, and the number of sentences that are rated class label $'0'$ is 174, whereas the important sentences that are related to class label $'1'$ to cover a summary of 400 words are 14 sentences, and the number of sentences that are rated class label '0' unimportant in the field of this

research is 168, leaving a large gap. As a result, the categorization and learning processes are ineffective and may produce erroneous findings, which is a concern. The dataset must be balanced as much as possible to tackle this problem. The Random Over Sampling (ROS) method, which helps increase the learning process and expects essential sentences that match the golden summary sentences written by experts, is one of the most well-known ways to accomplish balance. This method attempts to reduce variation as much as possible by randomly increasing the number of important sentences by applying equation (3.1). As a result, the number of important sentences increased to 168 to cover the summary of 200 words, and the number of unimportant sentences stayed at 174, while the number of important sentences increased to 168 to cover the summary of 400 words, and the number of unimportant sentences stayed at 168. Sometimes. Sometimes such equality occurs between labels and sentences, but not always.

Tabel4. 5 Correct Predicate in ANN Before and After Using ROS Method With Summary Size 200 Words

Topic	Golden Summary of words	Method	Number of samples	No. of class label “0”	No. of class label “1”	Correct predicate class label “0”	Correct predicate class label “1”
064	200	Before ROS Method	55	54	1	0.98	0.00
		After ROS Method	103	52	51	0.79	1.00
78	200	Before ROS Method	79	78	1	0.92	0.00

		After ROS Method	144	83	61	0.88	1.00
87	200	Before ROS Method	80	77	3	0.99	0.00
		After ROS Method	149	78	71	0.95	1.00
90	200	Before ROS Method	70	68	2	0.94	0.00
		After ROS Method	126	73	53	0.93	1.00
105	200	Before ROS Method	71	70	1	0.99	0.00
		After ROS Method	132	71	61	0.89	1.00
119	200	Before ROS Method	57	56	1	0.95	0.00
		After ROS Method	103	57	46	1.0	1.0

Table 4.5 demonstrates that before employing the ROS technique, the right percentage of the class label '1' is very bad. As a result, the ROS job is obvious and effective in terms of enhancing outcomes.

Tabel4. 6 Correct Predicate in Ann Before and After Using Ros Method with Summary Size 400 Words

Topic	Golden Summary of words	Method	Number of samples	No.of class label "0"	No.of class label "1"	Correct predicate label class "0"	Correct predicate label class "1"
064	400	Before ROS Method	55	51	4	0.92	0.25
		After ROS Method	101	55	46	0.84	1.00
078	400	Before ROS Method	79	75	4	0.96	0.00
		After ROS Method	146	72	74	0.92	1.00
87	400	Before ROS Method	80	74	6	0.99	0.33
		After ROS Method	147	73	74	0.97	1.00
090	400	Before ROS Method	70	68	2	0.97	0.50
		After ROS Method	133	62	71	0.97	0.93

105	400	Before ROS Method	71	68	1	0.97	0.00
		After ROS Method	134	55	79	0.98	1.00
119	400	Before ROS Method	57	55	2	0.95	0.00
		After ROS Method	105	47	58	0.91	1.00

Above, in Table 4.6, although the correct prediction percentage for the class label "0" before using the ROS method is perfect, the correct predicate percentage for the class label "1" is frustrating and bad. Except for topics $064,087, and$ 090, the class label "1" was achieved $25\%, 33\%, and$ 50% sequentially in the testing stage. This is due to unbalanced data. The results showed that the number of rows labeled "0" was greater than the number of rows labeled "1". This caused a failure in learning the backpropagation neural network correctly, and the training and testing stages became a failure. Therefore, using ROS, helped in increasing the correct predicate percentage class label "1", which represents the sentences matching the golden summary sentences, and it achieved high perfection. Therefore, the correct predicate percentage for the class label "1" in all topics became 1.00 for the same topics in the special, except for topic 090, which was improved to 0.93. This result is also considered perfect when compared with the above percentage for the same topic in particular.

4.2.3 New Summary Evaluation

This work deal with a text dataset named DUC 2002. The Document Understanding Conference (DUC) is the most common benchmarking dataset used for text summarization [101]. The DUC 2002 contains a set of topics up the number to 59 topics T= $\{t_1, t_2, \dots., t_n\}$, each topic includes a group of documents (articles) $t_i = \{d_1, d_2, \dots., d_k\}$ talking about that topic. each d contains a set of sentences $S = \{s_1, s_2, \dots., s_m\}$. All documents sentence special to a specific topic makes into one file $D^* = \{s_1, s_2, s_3, \dots., s_n\}$ for simplifying.

This section presents an evaluation of the new summary produced by the proposed summarizing system utilizing ROUGE metrics. In the context of summarization, the Rouge measures are the most typically utilized. This research focuses on three measures for comparing the candidate summary to the golden summary: $\text{Rouge} - 1, \text{Rouge} - 2$, and $\text{Rouge} - \text{L}$. In section 2.9, rouge metrics are thoroughly explained.

Two different volumes of golden summaries were taken for each topic and compared with the candidate summaries. One consists of 400 words and the other of 200 words. Table 4.7 describes number of sentences in all documents, and number of sentences in golden summary at 200 and 400 words for each topic of six topics taken in this work . On the other hand, the new candidate summaries are subject to several conditions before the evaluation process are: the first is a process of selecting the important sentences that bear the label "1" only; the second is checking the length candidate summary whether taller than golden summary, in this case, must remove some of the sentences in candidate summary which have the smallest score for length feature, and the final condition is the process of arranging (reordering) these sentences according to importance their depending on the date and time as additional features, and depending on centrality feature also in special cases, they will be detailed later. The dataset used in this dissertation contains articles of news, and thus the date and time become important features in this scope. They become a sequent of events accepted and understood during sentence candidate summary to achieve cohesive summary and consistency. Thus, reordering is considered more complex because it seeks to reorder sentences as a good way to simulate corresponding with the brain.

Table 4. 7 Displaying the Number of Sentences for Each Topic Compare With The Golden Summary Number Of Sentences for it in DUC 2002

Topic	No. of sentences in all documents for each topic	No. of sentences in the golden summary at 200 words	No. of sentences in the golden summary at 400 words
064	182	9	16
078	262	8	18
087	264	10	16
090	233	6	14
105	235	8	13
119	187	6	11

However, various issues arose during the implementation process. The first issue with extracting date and time features is that certain documents lack either the date or the time. Sometimes, only one of them is present. This problem was overcome by adding the word "$empt$" to refer to the date and the number "2400" to denote clock time because most documents only reach "2300" at most times. Another issue arises when candidate sentences have similar dates and the sentences are from different documents on the same topic. This challenge is overcome by emphasizing time (the clock) in the sorting process between these sentences, only then attaching them in the final candidate summary to complete the format. The third issue is that some of the potential sentences have the same date and clock time in some circumstances. This challenge is handled by relying on one of the eight proposed features. So, after many tests in selecting the feature on which the sentences are organized, the centrality feature ultimately depended on rearranging these phrases and then attaching them to the final candidate summary.

The reordering when reviewing the results of the candidate summaries for the six selected topics compared to the golden summary as shown, starting from Figure 4.2 to 4.3, it is noted that there is some discrepancy in locations between the candidate and golden summary sentences. The reason is that the process of arranging the sentences in the candidate summary was subject to three criteria or conditions are:

1. The main criterion in the arrangement of sentences in general in this dissertation is depending on the date.

2 . The first secondary criterion: if there are two or more candidate sentences with similar dates, then these sentences with similar dates will be taken and arranged in ascending order depending on the time (clock), and then they are returned based on the new order to the final candidate abstract.

3 . Second secondary criterion: If there are two or more sentences in the final candidate’s summary, that have a similar date and time, then those sentences will be taken and their positions re-arranged depending on the central feature, then they will be returned depending on the new arrangement to the final candidate summary.

There is an important point that must be made about this study: the sentence positions in the golden summary do not always appear in the same positions in the final candidate summary. For example, the sentence in the first position in the golden summary does not always appear in the same position in the final candidate summary. This comes back to One of two reasons is that, first, all the candidate summary sentences were subjected to the criteria indicated above, or second, perhaps, the proposed system did not predict that sentence. In the end, it leads to the following: there will be a process of crawling in the positions of the candidate sentences, so that the sentence in the second sequence in the golden summary may appear in the first position in the candidate summary, for instance.

Therefore, this dissertation focuses, by results analysis, on four sides. First, analyzing evaluation results obtained from the candidate summary compared with the reference summary using ROUGE metrics to find the degree of accuracy, precision, and recall this side will discussion in Table 4.8 which explains the results of the topics taken under the scope in this study, and the Table 4.9 shows average the topics results in a dataset in general. Second, Table 4.11 displays comparison results between our proposed system and the model described in [1]. Third, The Tables from 4.10 to 4.12 show and discuss a comparison between the final candidate summary generated at the test stage with the golden summary sentences for each of the six topics above, when the golden summary content is 200 or 400 words based on the answers to the three questions following: How many sentences in the candidate summary are predicated by the proposed system compared to the golden summary at 200 and 400 words for the topics mentioned above; why were one or more candidate sentences in the final candidate summary placed in positions that may differ from the position of these sentences in the golden summary; and why was more than one sentence chosen from the same chain (cluster) at times?

fourth, determining the amount of cohesiveness during analysis by computing the average of the consecutive sentences that matching between the candidate summary and the golden summary.

Tabel4. 8 Results Evaluation Cohesive Candidate Summary Using the Rouge Metrics

Topic	NO. of words	Rouge-1			Rouge-2			Rouge-L			No. of word new summary
		Recall	P	F-score	Recall	P	F-score	Recall	P	F-score	
064	200	0.96	1.0	0.98	0.92	0.97	0.95	0.96	1.0	0.98	188
	400	0.92	1.0	0.96	0.89	0.98	0.93	0.92	1.0	0.96	362
078	200	0.61	0.99	0.76	0.57	0.97	0.72	0.61	0.99	0.76	128
	400	0.86	0.99	0.92	0.82	0.96	0.88	0.86	0.99	0.92	329
087	200	0.84	0.98	0.90	0.84	0.97	0.90	0.84	0.98	0.90	168
	400	0.92	0.99	0.95	0.91	0.97	0.94	0.92	0.99	0.95	371
090	200	0.71	0.99	0.83	0.64	0.97	0.77	0.71	0.99	0.83	123
	400	0.66	1.0	0.79	0.61	0.97	0.75	0.66	1.0	0.79	242

105	200	0.84	1.0	0.91	0.78	0.96	0.86	0.84	1.0	0.91	164
	400	0.78	1.0	0.88	0.70	0.98	0.82	0.78	1.0	0.88	261
119	200	0.87	0.99	0.93	0.82	0.96	0.89	0.87	0.99	0.93	181
	400	0.92	0.99	0.95	0.90	0.98	0.94	0.92	0.99	0.95	379

The results of the evaluation scales in most of the six topics are encouraging, as seen in the Table above. When compared to the golden summary content of $200\ or\ 400$ words, the $F - score$ findings in the $Rouge - 1$ measure obtained remarkable levels of accuracy. Except for $Topic$ 078, which scored 0.76 against 200 words, and $Topic$ 090, which scored 0.79 against 400 words, the recall score reveals that the number of essential phrases predicted by the proposed system is smaller than the number of golden summary sentences. The extent of matching between predictive and reference sentences is reflected in the recall degree. Thus, $topic$ 078 has a recall degree of 0.61, while $topic$ 090 has a recall degree of 0.66. For topics 078, and 090, the f-score degree on the rouge-2 scale is lower than the rest of the topics for the same reason above, and in addition to that, the recall score for topic 078 is 0.57, and for topic 090 it is 0.64 for 200 words and 0.61 for 400 words. This rouge measures the bigram overlap between the candidate's computer-generated summary and the reference summaries. Thus, the bigram overlap refers between the candidate and the golden sentences summary is accepted but is small compared to the rest of the topics.

Tabel4. 9 Average the Rouge Evaluations Measure for Topic Summarization System Proposed to All Topics in The Dataset

Words	Rouge-1			Rouge-2			Rouge-L		
	Recall	P	F-score	Recall	P	F-score	Recall	P	F-score
200	0.69	0.99	0.81	0.63	0.97	0.75	0.69	0.99	0.81
400	0.63	0.99	0.76	0.56	0.96	0.69	0.63	0.99	0.76
200&400	0.66	0.99	0.78	0.60	0.97	0.72	0.66	0.99	0.78

The Table 4.9 above shows three of the average final evaluations. The first is the related results candidate summary at 200 words, the second is the related results candidate summary at 400 words, and the last is the mean results of the candidate summary related $200\ and\ 400$ words together, for all topics in DUC 2002, as determined by the proposed system. the results of the $\mathrm{Rouge} - 1, \mathrm{Rouge} - 2,$ and $\mathrm{Rouge} - \mathrm{L}$ tests revealed that the candidate cohesive summary generated by the proposed system is generally good. Thus, the proposed system can achieve a predicting-to-important sentence ratio greater than 50%, as evidenced by the recall scores for all of the above measures. The different degrees of $F - score$ are determined by the type of measure utilized, which was discussed in section 2.9.

A comparison with other related models should be made to evaluate our proposed model. In this dissertation, the model described in [1] is used for comparison. This model formulates content coverage and redundancy reduction challenges in a novel way. The GA method was used in [1] for comparison fairness. The evaluation measures rouge2, and ROUGE-L are used to compare the two models. These evaluation criteria were calculated by comparing computer-generated summaries against human-generated summaries. The machine-generated summary is evaluated by comparing it to the reference summary provided by professionals and supported by the DUC2002 dataset for each topic. The suggested model and the model introduced in [1] were tested on ten topics from the DUC2002 dataset [d061j, d062j, d063j, d064j, d065j, d066j, d067f, d068f, d069f, d070f]. Table 4.10 presents detailed average scores for the 20 runs to the rouge-2 and rouge-L scores, which describe the procedure of each topic in the model [1], compared with the results obtained from the proposed system.

Tabel4. 10 Detailed Results for Rouge-2 and Rouge-L Score

Topic number	Number of words	Model in [1]		Propose Model	
		Rouge-2	Rouge-L	Rouge-2	Rouge-L
061j	200	0.306	0.554	0.907	0.93
062j	200	0.200	0.481	0.752	0.815
063j	200	0.275	0.528	0.785	0.754
064j	200	0.233	0.488	0.95	0.98
065j	200	0.182	0.457	0.909	0.931
066j	200	0.181	0.441	0.832	0.892
067f	200	0.260	0.529	0.874	0.912
068f	200	0.496	0.626	0.673	0.734
069f	200	0.232	0.476	0.890	0.932
070f	200	0.262	0.513	0.820	0.872

The Table above demonstrates that the suggested system outperforms the model proposed in [1]. This suggests that the proposed system can provide a good cohesive summary and preserve the key phrases indicating the primary theme of the document collection form more accurate than the model proposed in [1] while simultaneously removing irrelevant and redundant ones from the entire collection.

Tabel4. 11 Compares Between the Correct Candidate Summary Sentences Prediction with The Golden Summary Sentences at Test 200 Words in Topic 064

Topic 064 with 200
Golden summary
the next east European McDonald's is scheduled to be opened in Budapest, Hungary, by the end of this year, said Vesna Milosevic, another Genex official. The communist world gets its first McDonald's next week, and some people here are wondering whether its American hamburgers will be as popular as the local fast-food treat, pljeskavica. the next east European McDonald's is Moscow (ap). The world's largest version of the landmark American fast-food chain rang up 30,000 meals on 27 cash registers, breaking the opening-day record for McDonald's worldwide, officials said. Shenzhen, china (ap). McDonald's hamburgers, fries, and golden arches came to China on Monday when the fast-food chain opened its first restaurant in a nation famed for its distinctive cuisine.McDonald's hopes to open a restaurant in Beijing later.Moscow. opening the second McDonald's restaurant in Moscow, along with a 12-story office block, Mr. George Cohon, head of McDonald's Canada, could well be described as the Russian authorities' idea of a model investor. Seoul. seven years after the u.s. hamburger giant first tried to bring its big macs to South Korea, the golden arches were finally going up. South Korea may be one of McDonald's most promising foreign markets, business analysts say.

Candidate Summary
The communist world gets its first mcdonald's next week, and some people here are wondering whether its merican hamburgers will be as popular as the local fast-food treat, pljeskavica. The next east european mcdonald's is scheduled to be opened in budapest, hungary, by the end of this year, said vesna milosevic, another genex official. South korea may be one of mcdonald's most promising foreign markets, business analysts say. Seven years after the u.s. hamburger giant first tried to bring its big macs to south korea, the golden arches were finally going up. The world's largest version of the landmark american fast-food chain rang up 30,000 meals on 27 cash registers, breaking the opening-day record for mcdonald's worldwide, officials said. Mcdonald's hamburgers, fries and golden arches came to china on monday when the fast-food chain opened its first restaurant in a nation famed for its distinctive cuisine. Mcdonald's hopes to open a restaurant in beijing later. Opening the second mcdonald's restaurant in moscow, along with a 12-storey office block, mr george cohon, head of mcdonald's canada, could well be described as the russian authorities' idea of a model investor.

Cluster name	Sentences candidate	date	Time
cluster0	the communist world gets its first mcdonald's next week, and some people here are wondering whether its merican hamburgers will be as popular as the local fast-food treat, pljeskavica.	14-Mar-88	1223
cluster1	the next east european mcdonald's is scheduled to be opened in budapest, hungary, by the end of this year, said vesna milosevic, another genex official.	14-Mar-88	1223
cluster1	south korea may be one of mcdonald's most promising foreign markets, business analysts say.	12-Apr-88	2400
cluster0	seven years after the u.s. hamburger giant first tried to bring its big macs to south korea, the golden arches were finally going up.	12-Apr-88	2400
cluster9	the world's largest version of the landmark american fast-food chain rang up 30,000 meals on 27 cash registers, breaking the opening-day record for mcdonald's worldwide, officials said.	31-Jan-90	2051
cluster4	mcdonald's hamburgers, fries and golden arches came to china on monday when the fast-food chain opened its first restaurant in a nation famed for its distinctive cuisine.	8-Oct-90	2027
cluster4	mcdonald's hopes to open a restaurant in beijing later.	8-Oct-90	2027
cluster7	opening the second mcdonald's restaurant in moscow, along with a 12-storey office block, mr george cohon, head of mcdonald's canada, could well be described as the russian authorities' idea of a model investor.	2-Jan-93	2400

Table 4.11 shows the number of sentences in the candidate summary as predicated by the proposed system, seven sentences, compared to the golden summary, nine sentences, at 200 words, for topic 064, which forms 80% of the number of sentences in the golden summary. The difference among sentence positions is that, for example, sentence number one is located in the first position, and sentence number two is located in the second position in the golden summary, while sentence number one becomes in the second position and the other one in the first position, respectively. Because the candidate sentences are subject to one of the three reordering criteria mentioned above, these two sentences are rearranged based on the centrality feature score because the date and time in these sentences are similar; thus, the sentence with a higher score in their document is introduced on the other. Also, note that sentences three and nine in the reference summary changed positions in the final summary since the reordering process in this state depends on old date priority because this dissertation deals with news articles, thus the date is considered very important. Therefore, sentence number three became in the fifth position because its date is "12-Apr-88", and sentence number nine became in the third position in the final summary because its date is "2-Jan-93". For easy understanding, Figure 4.2 displays a simple summary of the variance of the positions between the final and golden summary. Last, note that chosen two sentences from the same cluster (cluster) as in cluster 0, cluster 1, and cluster 4 because these sentences may be shared with one or more than the similarity of the words but not match them with the semantic, like sentence numbers 2, and 3 shares with the word "McDonald's"; while, sentences 6, and 7 shares with the three words "McDonald's", "open," and "restaurant" in the final summary; or maybe it is an important sentence, but it is an outlier, it does not have relation to sentences in those clusters, and during the reduction sentence procedure, it shows that it has a high score, like sentence 1, and 4 in cluster 0.

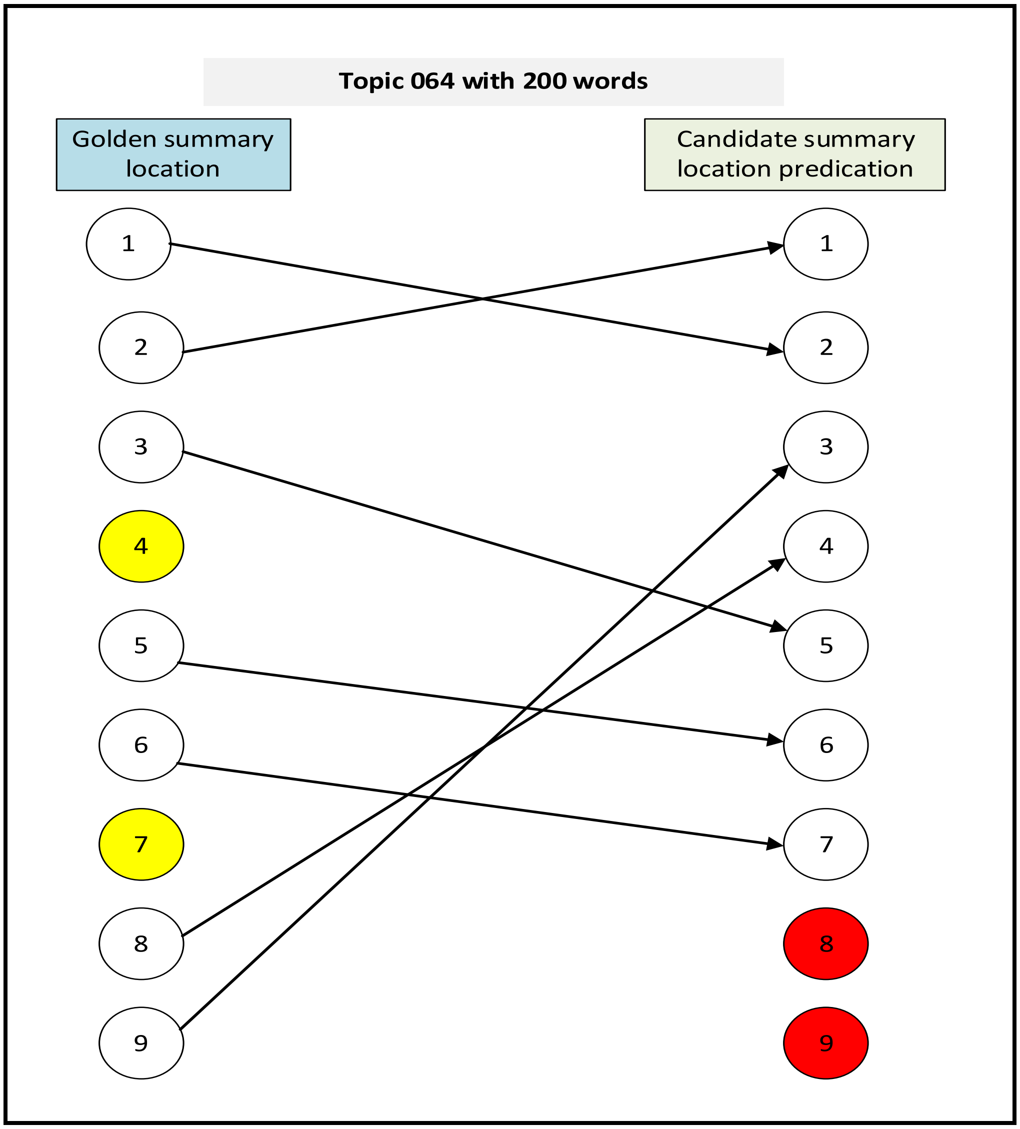

Figure 4. 2 Compares Candidate Sentence Positions with Original Sentence Positions in Golden Summary Special to Topic 064 At 200 Words

Figure 4.2 was referred to above because it helps with ease of understanding. In addition, through it, it is possible to calculate the percentage of the number of consecutive sentences that came in the final summary and match it with the golden summary to show the proportion of coherence and similarity between the final summary and the golden summary. Note that sentences No. 5 and 6 are consecutive sentences that appeared in the final summary in the same format in the golden summary, but in a different location for reasons discussed in Table 4.10 In addition, the unpredictability of the proposed system for the sentences 4 and 7 in the golden Abstract, led to the occurrence of the crawling process described earlier. Therefore, when calculating the ratio of successive sentences to the total number of sentences in the final summary, it results in 20% of the sentences being consecutive sentences similar to the golden summary in order in this topic at comparing with 200 words, regardless of the difference in locations between the two abstracts.

Tabel4. 12 Compares Between the Correct Candidate Summary Sentences Prediction with The Golden Summary Sentences at Test 400 Words in Topic 064

Topic 064 with 400
Golden Summary
belgrade, yugoslavia (ap).the communist world gets its first mcdonald's next week, and some people here are wondering whether its american hamburgers will be as popular as the local fast-food treat, pljeskavica.police kept watch on the lines of customers snaking around the block, and they regulated the number who came inside to avoid overcrowding.``i think this restaurant has no competition in belgrade,'' said milica danic, a housewife who treated her son to a cheeseburger.the next east european mcdonald's is scheduled to be opened in budapest, hungary, by the end of this year, said vesna milosevic, another genex official.moscow (ap).the world's largest version of the landmark american fast-food chain rang up 30,000 meals on 27 cash registers, breaking the opening-day record for mcdonald's worldwide, officials said.the restaurant, built by the company in a joint venture with the city of moscow that began 14 years ago, brought to 52 the number of countries where mcdonald's operates.mcdonald's built its own factory, including bakery, dairy, meat-processing plant and even potato storage yard, to provide its own guaranteed supplies in a country where up to 25 percent of the harvest rots en route to the consumer.shenzhen, china (ap).mcdonald's hamburgers, fries and golden arches came to china on monday when the fast-food chain opened its first restaurant in a nation famed for its distinctive cuisine.mcdonald's hopes to open a restaurant in beijing later.moscow.opening the second mcdonald's restaurant in moscow, along with a 12-storey office block, mr george cohon, head of mcdonald's canada, could well be described as the russian authorities' idea of a model investor.although the first moscow restaurant is the busiest mcdonald's in the world the plunging rouble, high inflation and constantly changing tax rules have been a challenge to mcdonald's management, who admitted that they have no idea of when it could recoup its initial dollars 50m investment.seoul seven years after the u.s. hamburger giant first tried to bring its big macs to south korea, the golden arches were finally going up.the company is a 50-50 joint venture of mcdonald's corp., oak brook, ill., and an accountant-turned-entrepreneur, ahn hyo young.south korea may be one of mcdonald's most promising foreign markets, business analysts say.the competition, besides wendy's international's 11 shops and six of burger king, a unit of pillsbury co., includes home-grown burger chains.

Candidte Summary
The communist world gets its first mcdonald's next week, and some people here are wondering whether its american hamburgers will be as popular as the local fast-food treat, pljeskavica. The next east european mcdonald's is scheduled to be opened in budapest, hungary, by the end of this year, said vesna milosevic, another genex official. ``i think this restaurant has no competition in belgrade,'' said milica danic, a housewife who treated her son to a cheeseburger. Police kept watch on the lines of customers snaking around the block, and they regulated the number who came inside to avoid overcrowding. South korea may be one of mcdonald's most promising foreign markets, business analysts say. The competition, besides wendy's international's 11 shops and six of burger king, a unit of pillsbury co., includes home-grown burger chains. The company is a 50-50 joint venture of mcdonald's corp., oak brook, ill., and an accountant-turned-entrepreneur, ahn hyo young. The world's largest version of the landmark american fast-food chain rang up 30,000 meals on 27 cash registers, breaking the opening-day record for mcdonald's worldwide, officials said. The restaurant, built by the company in a joint venture with the city of moscow that began 14 years ago, brought to 52 the number of countries where mcdonald's operates. Mcdonald's built its own factory, including bakery, dairy, meat-processing plant and even potato storage yard, to provide its own guaranteed supplies in a country where up to 25 percent of the harvest rots en route to the consumer. Mcdonald's hamburgers, fries and golden arches came to china on monday when the fast-food chain opened its first restaurant in a nation famed for its distinctive cuisine. Mcdonald's hopes to open a restaurant in beijing later. Opening the second mcdonald's restaurant in moscow, along with a 12-storey office block, mr george cohon, head of mcdonald's canada, could well be described as the russian authorities' idea of a model investor. Although the first moscow restaurant is the busiest mcdonald's in the world the plunging rouble, high inflation and constantly changing tax rules have been a challenge to mcdonald's management, who admitted that they have no idea of when it could recoup its initial dollars 50m investment.

Cluster Name	Sentences Candidate	Date	Time
cluster0	the communist world gets its first mcdonald's next week, and some people here are wondering whether its american hamburgers will be as popular as the local fast-food treat, pljeskavica.	14-Mar-88	1223
cluster1	the next east european mcdonald's is scheduled to be opened in budapest, hungary, by the end of this year, said vesna milosevic, another genex official.	14-Mar-88	1223
cluster5	``i think this restaurant has no competition in belgrade," said milica danic, a housewife who treated her son to a cheeseburger.	24-Mar-88	1514
cluster10	police kept watch on the lines of customers snaking around the block, and they regulated the number who came inside to avoid overcrowding.	24-Mar-88	1514
cluster1	south korea may be one of mcdonald's most promising foreign markets, business analysts say.	12-Apr-88	2400
cluster9	the competition, besides wendy's international's 11 shops and six of burger king, a unit of pillsbury co., includes home-grown burger chains.	12-Apr-88	2400

cluster0	the company is a 50-50 joint venture of mcdonald's corp., oak brook, ill., and an accountant-turned-entrepreneur, ahn hyo young.	12-Apr-88	2400
cluster9	the world's largest version of the landmark american fast-food chain rang up 30,000 meals on 27 cash registers, breaking the opening-day record for mcdonald's worldwide, officials said.	31-Jan-90	2051
cluster7	the restaurant, built by the company in a joint venture with the city of moscow that began 14 years ago, brought to 52 the number of countries where mcdonald's operates.	31-Jan-90	2051
cluster9	mcdonald's built its own factory, including bakery, dairy, meat-processing plant and even potato storage yard, to provide its own guaranteed supplies in a country where up to 25 percent of the harvest rots en route to the consumer.	31-Jan-90	2051
cluster4	mcdonald's hamburgers, fries and golden arches came to china on monday when the fast-food chain opened its first restaurant in a nation famed for its distinctive cuisine.	8-Oct-90	2027
cluster4	mcdonald's hopes to open a restaurant in beijing later.	8-Oct-90	2027
cluster7	opening the second mcdonald's restaurant in moscow, along with a 12-storey office block, mr george cohon, head of mcdonald's canada, could well be described as the russian authorities' idea of a model investor.	2-Jan-93	2400
cluster2	although the first moscow restaurant is the busiest mcdonald's in the world the plunging rouble, high inflation and constantly changing tax rules have been a challenge to mcdonald's management, who admitted that they have no idea of when it could recoup its initial dollars 50m investment.	2-Jan-93	2400

Table 4.12 shows the number of sentences in the candidate summary as predicted by the proposed system, fourteen sentences, compared to the golden summary, sixteen sentences, at 400 words, for topic 064, which forms 80% of the number of sentences in the golden summary. Note, that the first row is the golden summary, and the rest rows represent the candidate summary. The difference among sentence positions is that, for example, sentence number two is located in the second position, and sentence number three is located in the third position in the golden summary, while sentence number two becomes in the first position due to it has an older date which is “14-Mar-88” and the other one in the fourth position due to rearranged based on the centrality feature score because the date and time in between sentence number 3 and 4 sentences are similar. these two sentences are rearranged based on the centrality feature score because the date and time in these

sentences are similar; Also, note that all sentences in the reference summary changed positions in the final summary since the reordering process in this state depends on the old date priority because this dissertation deals with news articles, thus the date is considered very important. Therefore, all the candidate sentences are subject to one of the three reordering criteria mentioned above. For easy understanding, Figure 4.3 displays a simple summary of the variance of the positions between the final and golden summary. Last, note that chosen two sentences from the same cluster (cluster) as in cluster 0, cluster 1, cluster 4, cluster 7, and cluster 9 because these sentences may be shared with one or more than the similarity of the words but not match them with the semantic, like sentence numbers 11, and 12 shares with the word "McDonald's"; while, sentences 9, and 13 shares with the three words "McDonald's", "Moscow", and "restaurant" in the final summary; or maybe it is an important sentence, but it is an outlier, it does not have relation to sentences in those clusters, and during the reduction sentence procedure, it shows that it has a high score, like sentence 1, and 7 in cluster 0.

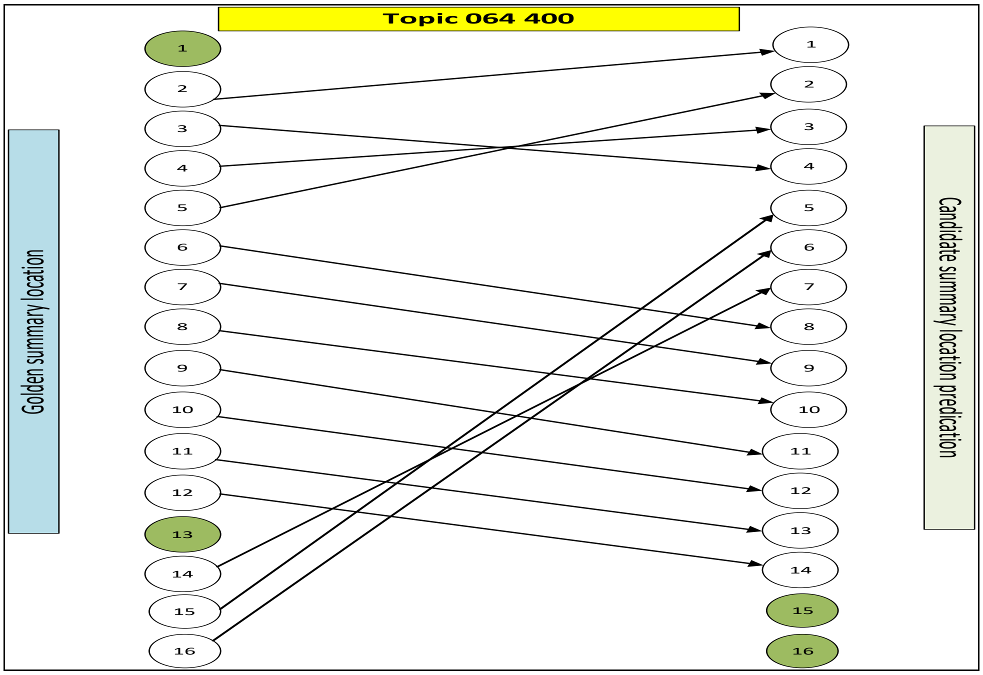

Figure 4. 3 Compares Candidate Sentence Positions with Original Sentence Positions In Golden Summary Special To Topic 064 At 400 Words

Figure 4.3 was referred to above because it helps with ease of understanding. In addition, through it, it is possible to calculate the percentage of the number of consecutive sentences that came in the final summary and match it with the golden summary to show the proportion of cohesion and similarity between the final summary and the golden summary. Note that sentences number $6,7,8,9,10,11, and\ 12$ are consecutive sentences that appeared in the final summary in the same format in the golden summary, but in a different location for reasons discussed in Table 4.12 In addition, the unpredictability of the proposed system for the sentences $1\ and\ 13$ in the golden Abstract, led to the occurrence of the crawling process described earlier. Therefore, when calculating the ratio of successive sentences to the total number of sentences in the final summary, it results in 50% of the sentences being consecutive sentences similar to the golden summary in order in this topic at comparing with 400 words for topic 064, regardless of the difference in locations between the two abstracts.

Also, this topic was extracted other golden summary consist of 22 sentences which represent 504 words by using QuillBot software to ensure the efficiency of this system's performance and compare matching with candidate summary as shown in Table 4.13. The Table 4.14 displays percentage of candidate summary sentences matching sentences in golden summary. the predicate candidate summary reach to 11 of sentences which mean 50%. Table 4.15 shows that the Results Evaluation Using the Rouge Metrics between cohesive candidate summary and golden summary which produced from QuillBot software proven that candidate summary achieved Encouraging and convincing results. The details of the rest of the topics taken as samples in this study are included in Appendix No. 1 at the end of the dissertation.

Tabel4. 13 Compares Between the Correct Candidate Summary Sentences Prediction with The Golden Summary Sentences which produced by QuillBot software at Test 504 Words in Topic 064

Golden Summary
Pljeskavica is made of ground pork and onions, and it is served on bread and eaten with the hands. The next East European McDonald's is scheduled to be opened in Budapest, Hungary, by the end of thisyear, said Vesna Milosevic, another Genex official. Negotiations have been going on for years for expanding the fast-food chain to the Soviet Union, but no agreement has been announced. The Belgrade media have suggested that the success of McDonald's in Yugoslavia depends on its acceptance by citizens long accustomed to a hamburger-like fast-food dish called the Pljeskavica: ground pork and onions on a bun. " The Big Mac meal, consisting of a hamburger, soft drink and french fries costs the equivalent of $2.57, or about as much the similar meal would cost in numerous Pljeskavica joints around town. Sadik Seljami, a waiter in a small Pljeskavica outlet just a few hundred yards from the McDonald's, suggested that the American restaurant wants to drive Yugoslav fast-food outlets out of business. The American corporation plans to open five additional restaurants Yugoslavia in the next five years. The next East European McDonald's, and the first in a Soviet bloc country, is to open next month in Budapest, Hungary. The world's largest version of the landmark American fast-food chain rang up 30,000 meals on 27 cash registers, breaking the opening-day record for McDonald's worldwide, officials said. The crush of customers was so intense the company stayed open until midnight, two hours later than planned. The previous opening-day record for sales was in Budapest, company officials said. McDonald's of Canada Chairman George Cohon, the man behind the deal, said many people were buying multiple orders and the restaurant served 15,000 to 20,000 people in just the first five hours of operation. Hundreds of Chinese waited for hours outside the restaurant in Shenzhen, an economic boom town near Hong Kong, for their first taste of a McDonald's hamburger, fries or shake. It is estimated there are more than 5,000 different Chinese dishes. However, many Chinese, who earn an average $32 a month, are still unable to afford fancy meals at restaurants. Another U.S. fast food outlet, Kentucky Fried Chicken, opened a restaurant in Beijing in 1987, and it now has four outlets there. McDonald's hopes to open a restaurant in Beijing later. The 500-seat McDonald's restaurant in a three-story building is operated by McDonald's Restaurant Shenzhen Ltd., a wholly owned subsidiary of McDonald's Hong Kong. But McDonald's executives hope eventually to get their supplies of beef and potatoes from China. Daniel Ng, chairman of McDonald's Restaurant Shenzhen Ltd., said it took two years to prepare for the restaurant's opening and his ambition now is to serve 10 million burgers a year. A waiter at the restaurant can earn 53 cents an hour.
Candidate Summary predication
Negotiations have been going on for years for expanding the fast-food chain to the soviet union, but no agreement has been announced.Sadik seljami, a waiter in a small pljeskavica outlet just a few hundred yards from the mcdonald's, suggested that the american restaurant wants to drive yugoslav fast-food outlets out of business.The belgrade media have suggested that the success of mcdonald's in yugoslavia depends on its acceptance by citizens long accustomed to a hamburger-like fast-food dish called the pljeskavica: ground pork and onions on a bun. The big mac meal, consisting of a hamburger, soft drink and french fries costs the equivalent of $2.57, or about as much the similar meal would cost in numerous pljeskavica joints around town. The american corporation plans to open five additional restaurants yugoslavia in the next five years.The next east european mcdonald's, and the first in a soviet bloc country, is to open next month in budapest, hungary.The previous opening-day record for sales was in budapest, company officials said.Mcdonald's hopes to open a restaurant in beijing later. The 500-seat mcdonald's restaurant in a three-story building is operated by mcdonald's restaurant shenzhen

ltd., a wholly owned subsidiary of mcdonald's hong kong. Another u.s. fast food outlet, kentucky fried chicken, opened a restaurant in beijing in 1987, and it now has four outlets there. Daniel ng, chairman of mcdonald's restaurant shenzhen ltd., said it took two years to prepare for the restaurant's opening and his ambition now is to serve 10 million burgers a year. But mcdonald's executives hope eventually to get their supplies of beef and potatoes from china. It is estimated there are more than 5,000 different chinese dishes. A waiter at the restaurant can earn 53 cents an hour.

Tabel4. 14 Percentage of Candidate Summary Sentences Matching Sentences in Golden Summary

Topic	No. of words Golden summary	The percentage of correct candidate sentences
064	200	80%
	400	80%
	504	50%

Tabel4. 15 Results Evaluation Cohesive Candidate Summary Using the Rouge Metrics

Topic	NO. of words	Rouge-1			Rouge-2			Rouge-L			No. of word new summary
		Recall	P	F-score	Recall	P	F-score	Recall	P	F-score	
064	504	0.71	1.0	0.83	0.62	0.97	0.75	0.71	1.0	0.83	286

The results explained above prove that the proposed system can access the important sentences and form a candidate summary free from redundant sentences, thereby achieving diversity and coverage. In addition, the system was able to create a cohesive and understandable summary so that the percentage of the consecutive sentences is very close to the percentage of consecutive sentences in a golden summary. Also, it succeeded in building a lexical chain based on sentences (LCS), solving traditional lexical chain problems. It is worth noting the success fantastic of the Developed clustering algorithm compared with the K – means algorithm despite some minor failures. However, each algorithm has failures and successes according to the nature of the data applied to it, and on the other hand, its effective contribution to building a LCS.

4.2.4 Validation Of The Proposed System Performance

This dissertation collected another dataset by picked news articles on certain topics from various news agencies over the internet and extracted the golden summary for each topic using QuillBot software to ensure the efficiency of this system's performance. The Quillbot is an online program that includes text summaries and paraphrase tools for authors, students, and marketers [102]. Table 4.16 describes number of sentences for each topic with their golden summary

Tabel4. 16 Number of Sentences for Each Topic with Their Golden Summary

Topic	No. of sentences in all documents	No. of sentences in golden summary
A132	142	17
B151	90	3
C185	295	20

This dissertation focuses, by results analysis, on three sides. First, analyzing evaluation results obtained from the candidate summary compared with the reference summary using ROUGE metrics to find the degree of accuracy, precision, and recall this side will discussion in Table 4.17 which explains the results of the topics taken under the scope in this dataset, Second, The Table s from 4.18 to 4.20 show and discuss a comparison between the final candidate summary generated at the test stage with the golden summary sentences for each of the three topics above, when the golden summary content is 390,114, or 518 words based on the answers to the three questions following: How many sentences in the candidate summary are predicated by the proposed system compared to the golden summary at different sizes of the words for the topics mentioned above; why were one or more candidate sentences in the final candidate summary placed in positions that may differ from the position of these sentences in the golden summary; and why was more than one sentence chosen from the same chain (cluster) at times?

Third, determining the amount of cohesiveness during analysis by computing the average of the consecutive sentences that matching between the candidate summary and the golden summary, this side will discuss topics results taken as sample related in this dateset shown in the Figures which started from 4.4 to 4.6.

Tabel4. 17 Results Evaluation Cohesive Candidate Summary Using the Rouge Metrics

Topic	NO. of words	Rouge-1			Rouge-2			Rouge-L			No. of word new summary
		Recall	P	F-score	Recall	P	F-score	Recall	P	F-score	
A132	390	0.67	0.87	0.76	0.57	0.74	0.64	0.67	0.87	0.76	303
B151	114	0.55	0.93	0.69	0.48	0.86	0.62	0.55	0.93	0.69	63
C185	518	0.93	1.0	0.96	0.88	0.97	0.92	0.93	1.0	0.96	480

The results of the evaluation scales are fantastic in most of the three topics above, as seen in the Table above. When compared to the golden summary content with different sizes of words, the F-score findings in the Rouge-1 measure obtained remarkable levels of accuracy. Except for Topic B151, which scored 0.69 against $114\ words,$ the recall score reveals that the number of essential phrases predicted by the proposed system is smaller than the number of golden summary sentences because the number of sentences entered to proposed system in test stage are $45\,sentences$, and that the number of correct sentences existence in this sample are $2\ sentences$. Therefore, the recall degree is low and the f-score is low too compared rest topics. For topics $B151$ the $f - score$ degree on the $rouge - 2$ scale is lower than the rest of the topics for the same reason above. This rouge measures the bigram overlap between the candidate's computer-generated summary and the reference summaries. Thus, the bigram overlap refers between the candidate and the golden sentences summary is accepted but is small compared to the rest of the topics.

Tabel4. 18 Compares Between the Correct Candidate Summary Sentences Prediction with The Golden Summary Sentences at Test 390 Words in Topic A132

Golden Summary
rubin, on a tour of asia's economic trouble spots, arrived from beijing, where he had accompanied u.s. president bill clinton on a visit. rubin will leave monday for thailand and south korea.the thai government has shut down 56 finance companies and taken over four small banks to sort out its financial sector.nairobi, kenya (ap) _ u.s. treasury secretary robert rubinfriday urged kenya to combat corruption and get its economy back ontrack.he was scheduled to meet with east african central bankers andfinance ministers later friday before concluding his week-longafrican trip that also included stops in ivory coast, south africa,namibia and mozambique.new york _ robert rubin, before joining the clintonadministration, built his career at goldman, sachs & co., tradingstakes in companies as head of that wall street firm's arbitragedesk. rubin and stephen friedman, co-heads of goldman from 1990 to 1992, had thebiggest stakes in the goldman partnership in the early 1990's, whenthe firm's top officers still could hold as much as 5 percent ofthe partnership, according to people at goldman. (joining the clinton white house in january1993, rubin was appointed treasury secretary in december 1994. ms. smith said rubin ''has no economic stake in goldman's success or failure. ''robert rubin's decision to step aside as treasury secretary wasresponsibly timed in that the major crises in the world now are notprimarily economic. rubin, a former co-chairman of the investment bank goldman,sachs, was chairman of president clinton's national economiccouncil until he was chosen as treasury secretary in late 1994,shortly before the mexican crisis erupted as that country'scurrency fell sharply.lawrence summers, the deputy secretary named to succeed rubin, is an outstanding if sometimes undiplomatic economist who is well qualified and deserves prompt confirmation.and treasury secretary robert rubin, who announced hisresignation wednesday, has joined greenspan in the pantheon ofmoney gods. (he arrived in washington as a supposedchampion of the poor.though rubin has bucked wall street by opposing a capital-gains taxcut, he comes from wall street, and most of his closest friends andadvisers are wall streeters, and he generally heeds the street.during the global financial crisis, rubin has halfheartedlywarned americans not to invest cavalierly in weak foreigneconomies, but anytime foolhardy american investors have beenthreatened, rubin has rushed to save them.he insists he was not trying to help wall street.be nice to wall street, and perhaps youcan be a money god, too.
Candidate Summary
Rubin will leave monday for thailand and south korea. The thai government has shut down 56 finance companies and taken over four small banks to sort out its financial sector. Nairobi, kenya (ap) _ u.s. treasury secretary robert rubin friday urged kenya to combat corruption and get its economy back on track. He was scheduled to meet with east african central bankers and finance ministers later friday before concluding his week-long african trip that also included stops in ivory coast, south africa, namibia and mozambique. Rubin and stephen friedman, co-heads of goldman from 1990 to 1992, had the biggest stakes in the goldman partnership in the early 1990's, when the firm's top officers still could hold as much as 5 percent of the partnership, according to people at goldman. Rubin, a former co-chairman of the investment bank goldman, sachs, was chairman of president clinton's national economic council until he was chosen as treasury secretary in late 1994, shortly before the mexican crisis erupted as that country's currency fell sharply. Robert rubin's decision to step aside as treasury secretary was responsibly timed in that the major crises in the world now are not primarily economic. Though rubin has bucked wall street by opposing a capital-gains tax cut, he comes from wall street, and most of his closest friends and advisers are wall streeters, and he generally heeds the street. During the global financial crisis, rubin has halfheartedly warned americans not to invest cavalierly in weak foreign economies, but anytime foolhardy american investors have been threatened, rubin has rushed to save them. He insists he was not trying to help wall street. and treasury secretary robert rubin, who announced his resignation wednesday, has joined greenspan in the pantheon of money gods. be nice to wall street, and perhaps you can be a money god, too.

Cluster Name	Sentences Candidate	Date	Time
cluster0	rubin will leave monday for thailand and south korea.	28-Jun-98	536
cluster0	the thai government has shut down 56 finance companies and taken over four small banks to sort out its financial sector.	30-Jun-98	339
cluster2	nairobi, kenya (ap) _ u.s. treasury secretary robert rubin friday urged kenya to combat corruption and get its economy back on track.	17-Jul-98	732
cluster0	he was scheduled to meet with east african central bankers and finance ministers later friday before concluding his week-long african trip that also included stops in ivory coast, south africa, namibia and mozambique.	17-Jul-98	732
cluster1	rubin and stephen friedman, co-heads of goldman from 1990 to 1992, had the biggest stakes in the goldman partnership in the early 1990's, when the firm's top officers still could hold as much as 5 percent of the partnership, according to people at goldman.	10-Aug-98	2135
cluster2	rubin, a former co-chairman of the investment bank goldman, sachs, was chairman of president clinton's national economic council until he was chosen as treasury secretary in late 1994, shortly before the mexican crisis erupted as that country's currency fell sharply.	12-May-99	2159
cluster3	robert rubin's decision to step aside as treasury secretary was responsibly timed in that the major crises in the world now are not primarily economic.	12-May-99	2159
cluster1	though rubin has bucked wall street by opposing a capital-gains tax cut, he comes from wall street, and most of his closest friends and advisers are wall streeters, and he generally heeds the street.	13-May-99	1310
cluster3	during the global financial crisis, rubin has halfheartedly warned americans not to invest cavalierly in weak foreign economies, but anytime foolhardy american investors have been threatened, rubin has rushed to save them.	13-May-99	1310
cluster4	he insists he was not trying to help wall street.	13-May-99	1310
cluster1	and treasury secretary robert rubin, who announced his resignation wednesday, has joined greenspan in the pantheon of money gods.	13-May-99	1310
cluster1	be nice to wall street, and perhaps you can be a money god, too.	13-May-99	1310

Table 4.18 shows the first row is the golden summary, seventeen sentences, at 390 words, and the rest rows are the number of sentences in the candidate summary as predicted by the proposed system, twelve sentences, compared to the golden summary, for topic $A123$, which forms 70.5% of the number of sentences in the golden summary. Note, the first row is the golden summary, and the rest rows represent candidate summary. The difference among sentence positions is that, for example, sentence number two is located in the second position, and sentence number three is located in the third position in the

golden summary, while sentence number two becomes in the first position and the sentence number three became in the second position due to rearranged based on the date priority. Typically, all sentences in the reference summary changed positions in the final summary since the reordering process in this state depends on the old date priority because this dissertation deals with news articles, thus the date is considered very important. Therefore, all the candidate sentences are subject to one of the three reordering criteria mentioned above. For easy understanding, Figure 4.4 displays a simple summary of the variance of the positions between the final and golden summary. Last, note that chosen two sentences from the same cluster (cluster) as in cluster 0 because these sentences may be shared with one or more than the similarity of the words but not match them with the semantic, or maybe it is an important sentence, but it is an outlier, it does not have relation to sentences in those clusters, and during the reduction sentence procedure, it shows that it has a high score.

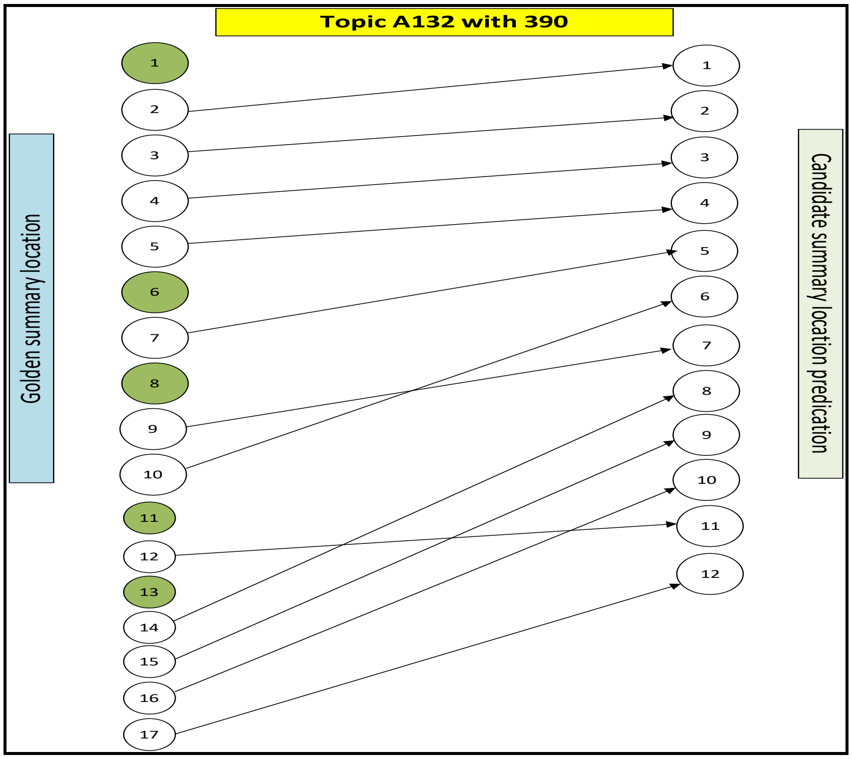

Figure 4. 4 Shows Candidate Sentence Positions Compared With Original Sentence Positions In Golden Summary Special To Topic A132 At 390 Words

In the preceding paragraph, Figure 4.4 was mentioned since it aids incomprehension. It is also possible to calculate the percentage of the number of consecutive sentences that appeared in the final summary and compare it to the golden summary to show the proportion of cohesion and similarity between the final summary and the golden summary using the information provided by this Figure. Keep in mind that sentences numbers $1, 2, 3, 4, 8, 9, 10, and\ 12$ are consecutive sentences that appeared in the final summary in the same structure as the golden summary, but in a different location for the reasons described in Table 4.18. Table 4.18: As previously said, the unpredictability of the suggested system for the sentences $1, 6, 8, 11, and\ 13$ in the Golden Abstract resulted in the occurrence of the crawling process, which was previously reported. Consequently, when the ratio of consecutive sentences to the total number of sentences in the final summary is calculated, it results in $66\ \%$ of the sentences in this topic being consecutive sentences similar to the golden summary in order, for topic $A132$, regardless of the difference in locations between the two abstracts. Typically, the most intriguing sentences in any article are found in the beginning sentences, and the final sentences in news articles are found in the last sentences. As a result, the suggested system in this field allows for the most precise prediction and rearrangement of significant sentences.

Tabel 4. 19 Compares Between the Correct Candidate Summary Sentences Prediction with The Golden Summary Sentences at Test 114 Words in Topic B151

greenspan, who announced a quarter-point interest rate hike tuesday, ends his current four-year term on june 30, 2000. republican elizabeth dole approves of greenspan's performance and would reappoint him if she were elected, said her spokesman ari fleischer. Arizona sen. john mccain approves of Greenspan, but thinks it is premature to consider his reappointment now, said spokesman howard opinsky. since Alan Greenspan haschaired the federal reserve, inflation and unemployment are both down nearly 2percentage points, interest rates are down 2.5 percentage points, the dow is up325 percent and the economy has grown at a 3.2 percent annual rate."

cluster4	Arizona sen. john mccain approves of greenspan, but thinks it is premature to consider his reappointment now, said spokesman howard opinsky.	24-Aug-99	1540
cluster0	since Alan Greenspan has chaired the federal reserve, inflation and unemployment are both down nearly 2 percentage points, interest rates are down 2.5 percentage points, the dow is up 325 percent and the economy has grown at a 3.2 percent annual rate."	4-Jan-00	2025

Table 4.19 shows the first row is the golden summary, three sentences, at 114 words, and the rest rows are the number of sentences in the candidate summary as predicted by the proposed system, two sentences, compared to the golden summary, for topic $B151$, which forms 66% of the number of sentences in the golden summary. because each sentence in the golden summary represents around 33.3%, therefore this system predicated two sentences from three origins. Note, the first row is the golden summary, and the rest rows represent candidate summary. The difference among sentence positions is that, for example, sentence number two is located in the second position, and sentence number three is located in the third position in the golden summary, while sentence number two becomes in the first position and the sentence number three became in the second position due to rearranged based on the date priority. Because the second sentence date is $4 - Jan - 00$ (the 00 here represents the year 2000). Typically, all sentences in the reference summary changed positions in the final summary since the reordering process in this state depends on the old date priority because this dissertation deals with news articles, thus the date is considered very important. Therefore, all the candidate sentences are subject to one of the three reordering criteria mentioned above. For easy understanding, Figure 4.5 displays a simple summary of the variance of the positions between the final and golden summary.

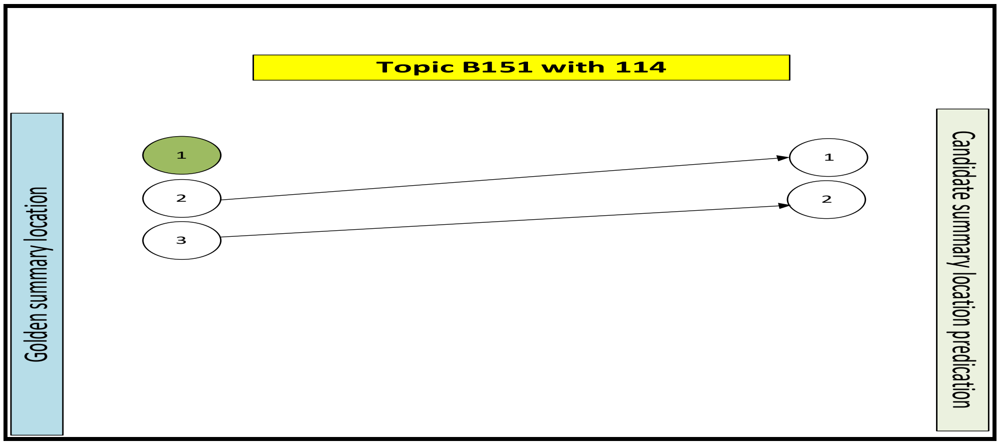

Figure 4. 5 Shows Candidate Sentence Positions Compared with Original Sentence Positions in Golden Summary Special to Topic B151 At 114 Words

In the preceding paragraph, Figure 4.5 was mentioned since it aids incomprehension. It is also possible to calculate the percentage of the number of consecutive sentences that appeared in the final summary and compare it to the golden summary to show the proportion of cohesion and similarity between the final summary and the golden summary using the information provided by this Figure. Keep in mind that sentences numbers $1, and\ 2$ are consecutive sentences that appeared in the final summary in the same structure as the golden summary, but in a different location for the reasons described in Table 4.19. Table 4.19: As previously said, the unpredictability of the suggested system for the sentence 1 in the Golden Abstract resulted in the occurrence of the crawling process, which was previously reported. Consequently, when the ratio of consecutive sentences to the total number of sentences in the final summary is calculated, it results in 100% of the sentences in this topic being consecutive sentences similar to the golden summary in order, for topic $B151$, regardless of the difference in locations between the two abstracts. Typically, the most intriguing sentences in any article are found in the beginning sentences, and the final sentences in news articles are found in the last sentences. As a result, the suggested system in this field allows for the most precise prediction and rearrangement of significant sentences.

Table 4. 20 Compares Between the Correct Candidate Summary Sentences Prediction with The Golden Summary Sentences at Test 677 Words in Topic C185

<table>
<tr><td colspan="4">the rev henry lyons says the biggest difference between himself -- the prominent leader of a large black church organization, the national baptist convention usa -- and president clinton is that ordinary americans already have forgiven clinton. if convicted, he could be sentenced to 30 years in prison. the case started in the summer of 1997 when lyons' wife was charged with setting fire to a $700,000 waterfront home he owned with another woman. lyons' claims that the convention had more than 8.5 million black members was a hoax, prosecutors said. lyons was to deliver black baptists and the convention would receive 33 percent of sales through commissions. instead, prosecutors say, just $187,600 of loewen's money went to market cemetery plots, while more than $1.6 million went to lyons, his friends and family. largo, fla. (ap) -- the rev henry lyons put a huge down payment on a $700,000 home, bought a 5.5-carat diamond ring and spent lavishly on cars, furniture and clothing, prosecutors say. lyons and his alleged mistress, bernice edwards, are accused of swindling more than $4 million from companies hoping to sell cemetery products, life insurance policies and credit cards to convention members. defense lawyers say failed business deals are not criminal matters. he testified the convention claimed more than 8 million members years before lyons became president. prosecutors have said an 8.5 million membership Figure was a hoax -- and lyons and ms. edwards used it to court companies seeking access to the group's members, promising them use of the convention's mailing list. largo, fla. (ap) -- after the rev henry lyons declined to testify in his racketeering trial, his co-defendant and alleged mistress told jurors she didn't advise lyons of her own criminal conviction when she began working for him. bernice edwards and lyons, president of the national baptist convention usa, are accused of swindling more than $4 million from corporations seeking to do business with the influential black church organization.without telling jurors she was convicted in 1993 of embezzlement, ms. edwards said thursday that she was put on probation for three years shortly before meeting lyons at a church function. ms. edwards said she did not tell lyons about the conviction when she began working to help his election as president, but did inform him of it before being hired as the convention's public relations director.ms. edwards, who has been described as lyons' girlfriend, denied there was any romantic involvement when he offered her the public relations job in 1994. prosecutors say the two used the money to buy an expensive waterfront home, luxury cars and diamond jewelry. lawyers for lyons said his decision not to testify was made partly because he faces a federal trial in april on charges of fraud, extortion and money laundering. the former president of the powerful national baptist convention usa was to return to court today to receive his sentence, expected to be three to seven years in prison.on feb. 27, jurors convicted the 57-year-old minister of bilking more than $4 million from companies wanting to sell cemetery products, life insurance policies and credit cards to convention members.</td></tr>
<tr><td>cluster2</td><td>instead, prosecutors say, just $187,600 of loewen's money went to market cemetery plots, while more than $1.6 million went to lyons, his friends and family.</td><td>24-Jan-99</td><td>0951</td></tr>
<tr><td>cluster11</td><td>the case started in the summer of 1997 when lyons' wife was charged with setting fire to a $700,000 waterfront home he owned with another woman.</td><td>24-Jan-99</td><td>0951</td></tr>
<tr><td>cluster8</td><td>lyons' claims that the convention had more than 8.5 million black members was a hoax, prosecutors said.</td><td>24-Jan-99</td><td>0951</td></tr>
</table>

cluster0	the rev henry lyons says the biggest difference between himself -- the prominent leader of a large black church organization, the national baptist convention usa -- and president clinton is that ordinary americans already have forgiven clinton.	24-Jan-99	0951
cluster8	lyons was to deliver black baptists and the convention would receive 33 percent of sales through commissions.	24-Jan-99	0951
cluster12	if convicted, he could be sentenced to 30 years in prison.	24-Jan-99	0951
cluster7	he testified the convention claimed more than 8 million members years before lyons became president.	11-Feb-99	0133
cluster6	lyons and his alleged mistress, bernice edwards, are accused of swindling more than $4 million from companies hoping to sell cemetery products, life insurance policies and credit cards to convention members.	11-Feb-99	0133
cluster8	prosecutors have said an 8.5 million membership Figure was a hoax -- and lyons and ms. edwards used it to court companies seeking access to the group's members, promising them use of the convention's mailing list.	11-Feb-99	0133
cluster7	largo, fla. (ap) -- the rev henry lyons put a huge down payment on a $700,000 home, bought a 5.5-carat diamond ring and spent lavishly on cars, furniture and clothing, prosecutors say.	11-Feb-99	0133
cluster9	defense lawyers say failed business deals are not criminal matters.	18-Feb-99	2353
cluster7	largo, fla. (ap) -- after the rev henry lyons declined to testify in his racketeering trial, his co-defendant and alleged mistress told jurors she didn't advise lyons of her own criminal conviction when she began working for him.	18-Feb-99	2353
cluster3	ms. edwards said she did not tell lyons about the conviction when she began working to help his election as president, but did inform him of it before being hired as the convention's public relations director.	18-Feb-99	2353
cluster0	without telling jurors she was convicted in 1993 of embezzlement, ms. edwards said thursday that she was put on probation for three years shortly before meeting lyons at a church function.	18-Feb-99	2353
cluster5	lawyers for lyons said his decision not to testify was made partly because he faces a federal trial in april on charges of fraud, extortion and money laundering.	31-Mar-99	253
cluster6	bernice edwards and lyons, president of the national baptist convention usa, are accused of swindling more than $4 million from corporations seeking to do business with the influential black church organization.	31-Mar-99	253

Table 4.20 shows the first row is the golden summary, twenty sentences, at 518 words, and the rest rows are the number of sentences in the candidate summary as predicted by

the proposed system, sixteen sentences, compared to the golden summary, for topic $C185$, which forms 80% of the number of sentences in the golden summary. the difference among sentence positions is that, for example, sentence number two is located in the second position, and sentence number three is located in the third position in the golden summary, while sentence number two becomes in the sixth position and the sentence number three became in the second position due to rearranged based on the date priority. Typically, all sentences in the reference summary changed positions in the final summary since the reordering process in this state depends on the old date priority because this dissertation deals with news articles, thus the date is considered very important. Therefore, all the candidate sentences are subject to one of the three reordering criteria mentioned above. For easy understanding, Figure 4.6 displays a simple summary of the variance of the positions between the final and golden summary. Last, note that chosen two sentences from the same cluster (cluster) as in cluster 0 because these sentences may be shared with one or more than the similarity of the words but not match them with the semantic, or maybe it is an important sentence, but it is an outlier, it does not have relation to sentences in those clusters, and during the reduction sentence procedure, it shows that it has a high score.

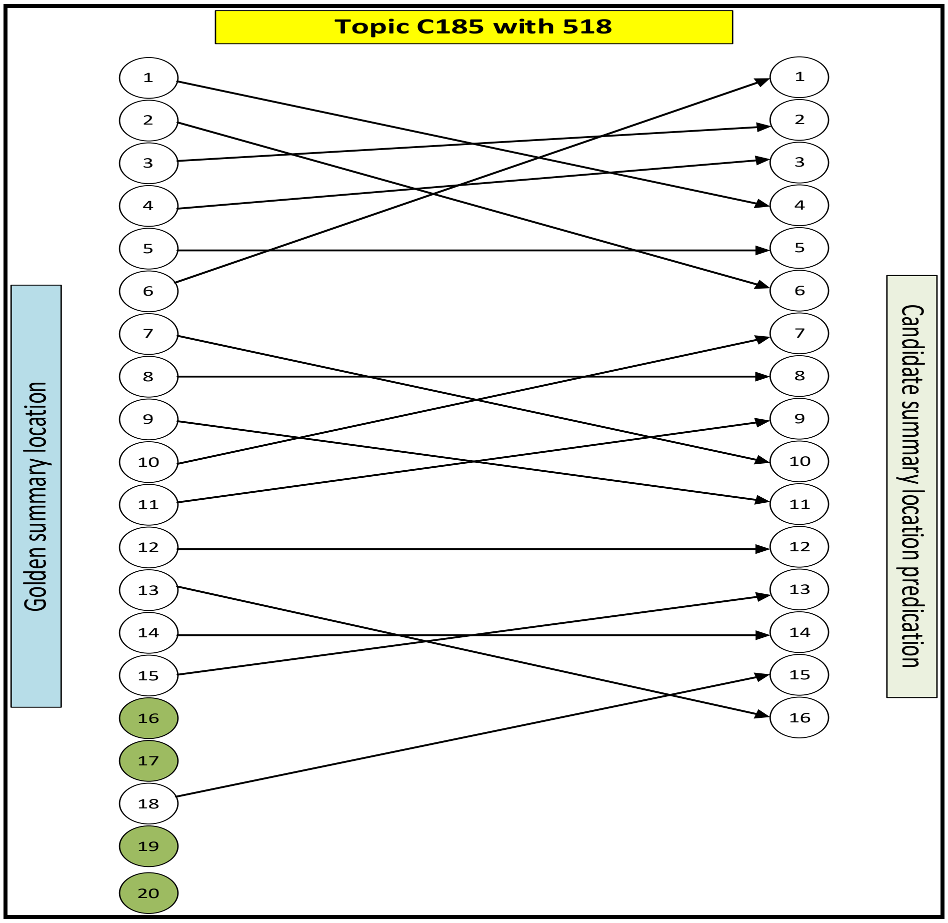

Figure 4. 6 Shows Candidate Sentence Positions Compared with Original Sentence Positions in Golden Summary Special to Topic C185 At 677 Words

In the preceding paragraph, Figure 4.6 was mentioned since it aids incomprehension. It is also possible to calculate the percentage of the number of consecutive sentences that appeared in the final summary and compare it to the golden summary to show the proportion of cohesion and similarity between the final summary and the golden summary using the information provided by this Figure. Keep in mind that sentences numbers $2, 3, 5,12, and\ 14$ are consecutive sentences that appeared in the final summary in the same structure as the golden summary, but in a different location for the reasons described in the Table 4.20. Table 4.20: As previously said, the unpredictability of the suggested system for the sentences $16,17,19, and\ 20$ in the Golden Abstract resulted in the occurrence of the crawling process, which was previously reported. Consequently, when the ratio of consecutive sentences to the total number of sentences in the final summary is calculated, it results in 31% of the sentences in this topic being consecutive sentences similar to the golden summary in order, for topic $c185,$ regardless of the difference in locations between the two abstracts, except location $5\ ,12, and\ 14$ same sentences are repeated in same position in both summarizes. Typically, the most intriguing sentences in any article are found in the first three sentences, and the final three sentences at least in news articles are found in the last sentences. As a result, the suggested system in this field allows for the most precise prediction and rearrangement of significant sentences.

Finally, the proposed system was able to predict the correct sentences matching with the golden summary in all topics as much as possible and succeeded in producing the summary arranged according to the conditions applied in the re-arranging process depended in this system. The difference in locations between the candidate’s summary and the golden one does not mean that the summary is incomprehensible and unclear, but this system has laid the foundations in the process to extract the important sentences and rearrange these sentences in an appropriate manner according to the sequence of time events that were referred to above. Therefore, through the candidate's summary, the reader can know the content of that topic.

CHAPTER FIVE

CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

With the Internet's exponential expansion and the availability of a massive amount of online information, identifying relevant content that satisfies user demands has become extremely difficult. This has sparked a race to develop technologies for an automatic document summarizing. Automated text summarization is regarded as an important technique in $Natural\ Language\ Processing$ (NLP) connected to text mining since it improves access to key sentences addressing the issue. The main advantage of text summarization is that the user's reading time can be minimized. The following can be conclusion based on the study findings and the proposed system details of:

- The proposed system is working based on generic classified.
- The Proposed System combines two concepts clustering and classification can be coined (Semi clustering unsupervised system based on extractive topic summarization).
- All sentences of documents are relevant to a specific topic gathered in one file for simplifying, then, similar semantic sentences will be collected from this file and put in an appropriate cluster coined-called chain sentence. After completing the assembly process, a lexical chain sentence (LCS) will be created.
- These proposed Developed clustering algorithm characteristics are different from clustering algorithms in principle because it does not require identifying the number of clusters at the start. but it decides the number of clusters automatically based on

the threshold value. while most cluster algorithms require identifying the number of clusters k in beginning like $Kmeans\ algorithm$.

- The collect numerical data are unlabeled in clusters. This is important for easily dealing. The clustering algorithms help to place close data in a specific cluster. This work has taken K- means clustering, and compared with to the proposed method results. The output of this approach clarified that it is the best in most experiments conducted. Also, this work success in creating a lexical chain based on sentence (LCS) thus become there flexible to deal with sentences as a complete sentence in the chain based on semantic sentence similarity.
- This dissertation succeeds in creating a lexical chain based on sentences (LCS), which is used to extract important sentences from them to form a summary.
- Through experiments, the results show that the use of the Random Over Sampling (ROS) method was very useful and prominent and helped in improving the results significantly. As for the proposed features, they helped in reaching the important sentences, which are characterized by readability and understanding. The class label for the created sentences is very useful in the process of classification by using the backpropagation neural network proposed model. The proposed model succeeds in classification right sentences, which handle class label '1' by giving it a high score. Focusing on the true positive (TP) idea in the confusion matrix to extract sentences matching the golden summary achieved impressive results. The reordering process succeeded based on the date feature and the other three considerations taken in this study in obtaining a cohesive and varied summary that covers the topic completely.
- Typically, In the future, this system could be applied to educational curricula to teach students who have difficulty in the learning process.

5.2 Future Work

The following points present some research directions for future work.

- The future work is to apply the developed clustering method to other data and compare it with other algorithms.
- Developing and training the proposed system to be capable of extracting the main ideas for various topics by developing and training it on larger datasets and storing it for the purpose of teaching it to the greatest number of datasets.
- Developing the performance of the proposed system by making it work on the principle of extraction and abstractive, which is called (the hybrid approach).
- Using different samples of data in different disciplines in the proposed algorithm to know the strengths and weaknesses in terms of its performance and how to develop it.
- Can developing the proposed system to work online in the future.
- Can use another algorithm intelligent like swarm optimization, deep learning, etc. to apply this work

References

[1] H. H. Saleh, N. J. Kadhim and B. A. Attea, "A Genetic Based Optimization Model for Extractive Multi-Document Text Summarization," *Iraqi Journal of Science,* pp. 1489-1498, 2015.

[2] R. Ahuja and W. Anand, "Multi-document Text Summarization," *Springer Nature Singapore Pte Ltd,* pp. 235-242, 2017.

[3] D. -Z. D. P. M. Pardalos and Z. Z. , "Nonlinear Combinatorial Optimization," *springer,* p. 295, 2010.

[4] M. A. S. P. M. A. S. S. E. D. Trippe, J. B. Gutierrez and K. K. , "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," *arXiv,* pp. 1-13, 2017.

[5] A. S. and S. M. , "Deep Learning in Text Summarization - A Survey," in *Alliance International Conference on Artificial Intelligence and Machine Learning (AICAAM),*, 2019.

[6] L. Lebanoff, K. Song and F. Liu, "Adapting the Neural Encoder-Decoder Framework from Single to Multi-Document Summarization," *arXiv,* vol. 2, 2018.

[7] R. M. Aliguliyev, " A new sentence similarity measure and sentence based extractive technique for automatic text summarization," *Elsevier-Expert Systems with Applications,* p. 7764 –7772, 2009.

[8] F. Bayatmakou, A. Mohebi and A. Ahmadi, "An interactive query-based approach for summarizing scientific documents," *Information Discovery and Delivery,* 2021.

[9] J. Thomas, A. Sreeraj, A. Sreeraj, M. M. Varghese and T. Kuriakose, "Automatic Text Summarization Using Deep Learning and Reinforcement Learning," *Sentimental Analysis and Deep Learning,* pp. 769-778, 2022.

[10] W. S. El-Kassas, C. R. Salama, A. A. Rafea and H. K. Mohamed, "Automatic Text Summarization: A Comprehensive Survey," *Expert Systems with Applications,* vol. 165, 2021.

[11] A. Nenkova and K. McKeown, "A Survey of Text Summarization Techniques," *Springer ,* pp. 43-76, 2012.

[12] V. Dalal and L. Malik, "A Survey of Extractive and Abstractive Text Summarization Techniques," in *6th International Conference on Emerging Trends in Engineering and Technology*, 2013.

[13] A. Tandel, B. Modi, P. Gupta, S. Wagle and S. Khedkar, "Multi-document text summarization - a survey," International Conference on Data Mining and Advanced Computing (SAPIENCE), 2016.

[14] M. Campr and K. Ježek , "Comparing Semantic Models for Evaluating Automatic Document Summarization," International Conference on Text, Speech, and Dialogue, 2015.

[15] L. Hou, P. Hu and C. Bei , "Abstractive Document Summarization via Neural Model with Joint Attention," National CCF Conference on Natural Language Processing and Chinese Computing, 2017.

[16] C. Mallick, A. K. Das, M. Dutta, A. K. Das and A. Sarkar, "Graph-Based Text Summarization Using Modified TextRank," *Springer Nature Singapore,* pp. 137-146, 2019.

[17] C. Mallick, M. Dutta, A. K. Das, A. Sarkar and A. K. Das, "Extractive Summarization of a Document Using LexicalChains," *Springer Nature Singapore Pte Ltd.,* pp. 825-836, 2019.

[18] I. K. Bhat, M. Mohd and R. Hashmy, "SumItUp: A Hybrid Single-Document Text Summarizer," *Springer,* pp. 619-634, 2018.

[19] M. Mohd, R. Jan and M. Shah, "Text Document Summarization using Word Embedding," *Expert Systems with Applications,* vol. 143, 2020.

[20] A. K. S. P. and R. P. , "Text Summarization from Legal Documents: A Survey," *Artificial Intelligence,* pp. 2-37, 2017.

[21] N. J. Kadhim and H. H. Saleh, "Evolutionary Based Extractive Multi-Document Text Summarization.," *thesis,* no. University of Technology. Department of Computer Science. , 2016.

[22] M. Afsharizadeh, H. E. Komleh and A. Bagheri, "Query-oriented Text Summarization using Sentence Extraction Technique," in *2018 4th International Conference on Web Research (ICWR)* , 2018.

[23] R. Ahuja and W. Anand, "Multi-document Text Summarization using sentence extraction," *Springer Nature Singapore Pte Ltd,* pp. 235-242, 2017.

[24] E. V. -Valdés, A. S. -Cuevas, J. A. Olivas and F. P. Romero, "A Fuzzy Approach for Sentences Relevance Assessment in Multi-document Summarization," *Springer Nature Switzerland,* pp. 57-67, 2019.

[25] C. Mallick, M. Dutta, A. K. Das, A. Sarkar and A. K. Das, "ExtractiveSummarization ofaDocumentUsingLexicalChains," *Springer Nature Singapore Pte Ltd. ,* pp. 825-836, 2019.

[26] H. Rezaei, S. A. Moeinzadeh, A. Shahgholian and M. Saraee, "Features In Extractive Supervised Single-Document Summarization: Case Of Persian News," *arXiv,* vol. 2, 2019.

[27] M. Zhong, P. Liu, Y. Chen, D. Wang, X. Qiu and X. Huang, "Extractive Summarization as Text Matching," *arxiv,* vol. 1, pp. 1-12, 2020.

[28] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east.," *IDC iView: IDC Analyze the future*, pp. 1-16, 2012.

[29] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani and M. R. Yeganegi, "Text mining in big data analytics," *Big Data and Cognitive Computing,* vol. 4, no. 1, pp. 1-34, 2020.

[30] Y. Kang, Z. Cai, C. W. Tan, Q. Huang and H. Liu, "Natural language processing (NLP) in management research: A literature review," *Journal of Management Analytics,* vol. 7, pp. 139-172, 2020.

[31] M. Berker and T. Güngör, "Using Genetic Algorithms with Lexical Chains for Automatic Text Summarization," in *In Proceedings of the 4th International Conference on Agents and Artificial Intelligence*, 2012.

[32] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM,* vol. 38, no. 11, pp. 39-41, 1995.

[33] A. Jain and A. Gaur, "Summarizing Long Historical Documents Using Significance and Utility Calculation Using Wordnet," *Imp. J. Interdiscip,* 2017.

[34] A. Pawar and V. Mago, "Calculating the similarity between words and sentences using a lexical database and corpus statistics," *arXiv preprint arXiv:1802.05667,* vol. 2, pp. 1-14, 2018.

[35] T. Wei, Y. Lu, H. Chang, Q. Zhoua and X. Bao, "A semantic approach for text clustering using WordNet and lexical chains," *Expert Systems with Applications,* vol. 42, pp. 2264 - 2275, 2015.

[36] B. R. Sinaga, A. and S. T. Widodo, "Cohesion and Coherence of Narrative Essays of Madagascar Students of Indonesian Language for Foreign Speakers (BIPA) at UPTP2B Sebelas Maret University," *Budapest International Research and Critics Institute-Journal (BIRCI-Journal),* 2019.

[37] A. Latifah and S. Triyono, "Cohesion and coherence of discourse in the story of "layangan putus" on social media facebook," *Indonesian Journal of EFL and Linguistics,* vol. 5, no. 1, pp. 1-16, 2020.

[38] M. Halliday and . R. Hasan, "Cohesion in English. London: Longman.," *The Oxford Essential Guide to Writing. New York: The Berkley,* 1976.

[39] A. Karadeniz, "Cohesion and Coherence in Written Texts of Students of Faculty of Education," *Journal of Education and Training Studies,* 2017.

[40] F. T. AL-Khawaldeh and V. W. Samawi, "Lexical Cohesion and Entailment based Segmentation for Arabic Text Summarization (LCEAS)," *World of Computer Science and Information Technology Journal (WCSIT),* 2015.

[41] H. Zheng and M. Lapata, "Sentence Centrality Revisited for Unsupervised Summarization," in *57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019.

[42] K. Nandhini and S. R. Balasundaram, "Significance of Learner Dependent Features for Improving Text Readability using Extractive Summarization," in *4th International Conference on Intelligent Human Computer Interaction,* Kharagpur, India, 2012.

[43] A. Conneau and K. Douwe , "Senteval: An evaluation toolkit for universal sentence representations," *arXiv*, vol. 1, pp. 1-6, 2018.

[44] D. Cer, Y. Yang, S. -yiKong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. G. -C´espedes, S. Yuan, C. Tar, Y. H. Sung, B. Strope and R. Kurzweil, "Universal Sentence Encoder," *arXiv,* 2018.

[45] T. Mikolov, S. Ilya , C. Kai , C. S. Greg and D. Jeff , "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems,* vol. 26, 2013.

[46] J. Pennington, S. Richard and M. D. Christopher , "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014.

[47] P. Koirala and N. B. Nobal, "Npvec1: Word embeddings for nepali-construction and evaluation," in *Proceedings of the 6th Workshop on Representation Learning for NLP*, 2021.

[48] M. Iyyer, V. Manjunatha, J. Boyd-Graber and H. D. III, "Deep unordered composition rivals syntactic methods for text classification," in *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing*, 2015.

[49] Y. Bengio, C. Aaron and V. Pascal , "A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 35, no. 8, p. 1798–1828, 2013.

[50] A. Martins and A. Ramon , "From softmax to sparsemax: A sparse model of attention and multi-label classification," in *International conference on machine learning.*, PMLR, 2016.

[51] A. Chaudhary, "Universal Sentence Encoder Visually Explained," Machine Learning Engineer at Fusemachines , 2022. [Online]. Available: https://amitness.com/2020/06/universal-sentence-encoder/.

[52] L. Chunjie, Z. Jianfeng, W. Lei and Y. Qiang, "Cosine normalization: Using cosine similarity instead of dot product in neural networks," in *International Conference on Artificial Neural Networks*, Cham, 2018.

[53] B. Li and L. Han, "Distance Weighted Cosine Similarity Measure for Text Classification," International conference on intelligent data engineering and automated learning, 2013.

[54] P. Xia, L. Zhang and F. Li, "Learning similarity with cosine similarity ensemble," *Information Sciences,* vol. 307, pp. 39-52, 2015.

[55] K. Park, J. S. Hong and W. Kim, "A Methodology Combining Cosine Similarity with Classifier for Text Classification," *Applied Artificial Intelligence,* vol. 5, pp. 396-411, 2020.

[56] S. Banerjee, P. Mitra and K. Sugiyama, "Multi-document abstractive summarization using ilp based multi-sentence compression," in *Twenty-Fourth International Joint Conference on Artificial Intelligence,* 2015.

[57] M. Campr and K. Jeˇzek, "Comparing Semantic Models for Evaluating Automatic Document Summarization," International Conference on Text, Speech, and Dialogue, 2015.

[58] K. Ganesan, "Rouge 2.0: Updated and improved measures for evaluation of summarization tasks," *arXiv,* 2018.

[59] S. Gupta and . S. Gupta, "Abstractive Summarization: An Overview of the State of the Art," *Expert Systems With Applications,* 2018.

[60] M. K. Chandrasekaran, M. Yasunaga, D. Radev, D. Freitag and M. Y. Kan, "Overview and Results: CL-SciSumm Shared Task 2019," *arxiv,* vol. 1, 2019.

[61] H. Arshad, "Learn ROUGE: Evaluation Metric for Text Summarization Task," ilmoirfan, COMSATS University Islamabad,, 2022.

[62] E. Rashedi , A. Mirzaei and . M. Rahmati, "An Information Theoretic Approach to Hierarchical Clustering Combination," *Neurocomputing,* vol. 148, p. 487–497, 2015.

[63] M. A. B. Siddique, R. B. Arif, M. M. Rahman Khan and Z. Ashrafi, "Implementation of Fuzzy C-Means and Possibilistic C-Means Clustering Algorithms, Cluster Tendency Analysis and Cluster Validation," *ArXiv,* 2018.

[64] H. Rezanková and E. B, "Cluster analysis and categorical data," *Statistika,* pp. 216-232, 2009.

[65] S. V. Wazarkar and A. A. Manjrekar, "Text Clustering Using HFRECCA and Rough K-Means Clustering Algorithm," *Discovery,* pp. 44-47, 2014.

[66] H. Aidos and A. Fred, "Statistical Modeling of Dissimilarity Increments for D-dimensional Data: Application in Partitional Clustering," *Elsevier,* vol. 45, no. 9, p. 3061–3071, 2012.

[67] W.-F. Hsiao and T.-M. Chang, "An Incremental Cluster-Based Approach to Spam Filtering," *Elsevier,* vol. 34, no. 3, p. 1599–1608, 2008.

[68] Z. Shi and L. S. Pun-Cheng, "Spatiotemporal Data Clustering: A Survey ofMethods," *ISPRS international journal of geo-information,* vol. 8, no. 2, pp. 112-128, 2019.

[69] P. BHATTACHARJEE and P. MITRA, "A survey of density based clustering algorithms," *Frontiers of Computer Science,* vol. 15, no. 1, pp. 1-27, 2021.

[70] F. d. A. de Carvalho, C. P. Tenório and . N. L. C. Junior, "Partitional Fuzzy Clustering Methods Based on Adaptive Quadratic Distances," *Elsevier,* vol. 157, no. 21, p. 2833 – 2857, 2006.

[71] A. Bouguettaya, . Q. Yu, . X. Liu, X. Zhou and . A. Song, "Efficient Agglomerative Hierarchical Clustering," *Elsevier,* vol. 42, no. 5, pp. 2785-2797, 2015.

[72] S. Liang, D. Han and Y. Yang, "Cluster validity index for irregular clustering results," *Applied Soft Computing Journal*, vol. 95, 2020.

[73] A. A. Abin and . H. Beigy, "Active Constrained Fuzzy Clustering: A Multiple Kernels Learning Approach," *Elsevier*, vol. 48, no. 3, pp. 953-967, 2015.

[74] J. C. Bezdek, Pattern recognition with fuzzy objective function algorithms, Springer Science & Business Media, 2013.

[75] T. Zhang, R. Raghu and L. Miron , "BIRCH: an efficient data clustering method for very large databases," in *In Proceedings of the ACM SIGMOD International Conference on Management of Data*, Canada, 1996.

[76] S. Guha, R. Rajeev and S. Kyuseok, "CURE: An efficient clustering algorithmfor large databases," in *In Proceedings of the ACM Sigmod record International Conference onManagement of Data*, USA, 1998.

[77] E. Masciari, M. M. Giuseppe and Z. Carlo , "A new, fast and accurate algorithm for hierarchical clustering on Euclidean distances," in *Pacific-asia conference on knowledge discovery and data mining*, Berlin, 2013.

[78] J. Rajaie and ,. Fakhar, "A Novel Method for Document Clustering using Ant-Fuzzy Algorithm," *The Journal of Mathematics and Computer Science*, vol. 4, no. 2, pp. 182 - 196, 2012.

[79] I. Chami, A. Gu, V. Chatziafratis and C. Ré, "From Trees to Continuous Embeddings and Back:Hyperbolic Hierarchical Clustering," *arXiv* , pp. 1-27, 2020.

[80] M. Ester, P. H. Kriegel, J. Sander and X. Xu, "A density-based algorithm for discovering clusters in large spatial," in *n Proceedings of the Proceedings of the 2nd International Conference on Knowledge and Discovery and Data Mining*, Portland,USA, 1996.

[81] M. Ankerst, M. M. Breunig, P. H. Kriegel and J. Sander, "OPTICS: Ordering points to identify the clustering structure," in *In Proceedings of the ACMSIGMOD Record*, Philadelphia,USA, 1999.

[82] H. Darong and P. Wang , "Grid-based DBSCAN algorithm with referential parameters," in *2012 International Conference on Applied Physics and Industrial Engineering*, 2012.

[83] M.-I. Akodj`enou-Jeannin,, K. Salamatian and P. Gallinari, "Flexible Grid-Based Clustering," In European Conference on Principles of Data Mining and Knowledge Discovery, 2007.

[84] G. H. Lee, "Grid-based dynamic clustering with grid," *Intelligent Data Analysis*, vol. 20, no. 4, pp. 853-875, 2016.

[85] P. A. Fattah, U. Aickelin and C. Wagner, "Clustering Human Behaviour in Public Good Experiments," Public Good Experiment, University of Nottingham, 2013.

[86] R. S. M. Patibandla and N. Veeranjaneyulu, "Performance analysis of partition and evolutionary clustering methods on various cluster validation criteria," *Arabian Journal for Science and Engineering,* vol. 43, no. 8, pp. 4379-4390, 2018.

[87] E. Rendón, I. Abundez, A. Arizmendi and E. M. Quiroz, "Internal versus External cluster validation indexes," *INTERNATIONAL JOURNAL OF COMPUTERS AND COMMUNICATIONS,* vol. 5, no. 1, 2011.

[88] D. L, DAVIES and D. W. BOULDIN, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vols. PAMI-1, no. 2, pp. 224-227, 1979.

[89] Z. Zheng, Y. Cai and Y. Li, "Oversampling Method For Imbalanced Classification," *Computing and Informatics,,* vol. 34, pp. 1017-1037, 2015.

[90] X. W. Liang, A. P. Jiang, T. Li, Y. Y. Xue and G. T. Wang, "Lr-Smote— An improved unbalanced data set oversampling based on K-means and SVM," *Elsevier-Knowledge-Based Systems,* vol. 196, pp. 1-10, 2020.

[91] A. F. Anifowose, L. Jane and A. Abdulazeez , "Hybrid intelligent systems in petroleum reservoir characterization and modeling: the journey so far and the challenges ahead," *Journal of Petroleum Exploration and Production Technology,* vol. 7, no. 1, pp. 251-263, 2017.

[92] B. Warsito, S. Rukun and Y. Hasbi , "Cascade forward neural network for time series prediction.," *In Journal of Physics: Conference Series,* vol. 1025, no. 1, p. 012097, 2018.

[93] P. Saikia, R. D. Baruah, S. K. Singh and P. K. Chaudhuri, "Artificial Neural Networks in the domain of reservoir characterization: A review from shallow to deep models," *Computers and Geosciences,* vol. 135, p. 104357, 2020.

[94] T. W. HUGHES, M. MINKOV, Y. SHI and S. FAN, "Training of photonic neural networks through in situ backpropagation and gradient measurement," *Optica,* vol. 5, no. 7, pp. 864-871, 2018.

[95] R. R. Asaad and R. I. Ali, "Back Propagation Neural networks(BPNN) and Sigmoid Activation Function in Multi-Layer Networks," *Academic Journal of Nawroz University (AJNU),* vol. 8, no. 4, p. 216, 2019.

[96] M. JANG and P. KANG, "Learning-Free Unsupervised Extractive Summarization Model," *IEEE,* vol. 9, pp. 14358-14368, 2021.

[97] A. D. Rasamoelina, F. Adjailia and P. Sinˇc´ak, "A Review of Activation Function for Artificial Neural Network," in *18th World Symposium on Applied Machine Intelligence and Informatics (SAMI),* 2020.

[98] C. E. Nwankpa, W. Ijomah, A. Gachagan and S. Marshall, "Activation Functions: Comparison of Trends in Practice and Research for Deep Learning," *arXiv,* vol. 1, 2018.

[99] NIST, "Document Understanding Conferences," oa Dang (hoa DOT dang AT nist.gov, [Online]. Available: https://duc.nist.gov/data.html.

[100] M. Afsharizadeh, H. E. Komleh and A. Bagheri, "Query-oriented Text Summarization using Sentence Extraction Technique," in *4th International Conference on Web Research (ICWR)*, 2018.

[101] N. Moratanch and S. Chitrakala, "A Survey on Abstractive Text Summarization," in *International Conference on Circuit, Power and Computing Technologies [ICCPCT]*, Nagercoil, India, 2016.

[102] 12 11 2021. [Online]. Available: https://thetechreviewer.com/quillbot-review/.

[103] V. PRADEEPIKA and H. OM, "A novel approach for text summarization using optimal combination of sentence scoring methods," *Springer*, 2019.

[104] M. A. B. Siddique, R. B. Arif, M. M. Rahman Khan and Z. Ashrafi, "Implementation of Fuzzy C-Means and Possibilistic C-Means Clustering Algorithms, Cluster Tendency Analysis and Cluster Validation," *ArXiv*, vol. 1, pp. 1-8, nov 2018.

[105] [Online]. Available: https://www.kaggle.com/datasets/gpreda/bbc-news.

APPENDIX 1

Table 1. Compares Between the Correct Candidate Summary Sentences Prediction with The Golden Summary Sentences at Test 200 Words In Topic 078

last year, the oscarcast was seen by one billion people in 79 countries.``ben-hur'' in 1959 was the most-awarded film with 11, and walt disney was the most-awarded person with 32 .what actress has received the most oscars?katharine hepburn, 4. The thalberg award has come in handy to acknowledge giants who were overlooked for individual awards.in a lecture and question-and- answer session monday, lee shrugged off the academy's icy treatment of his latest movie, ``do the right thing,'' and blamed it on generational politics.as it was, 10 different movies split 17 awards, the most winners for one year in this decade.given the tremendous impact both an oscar nomination and an oscar victory have on a film's fortunes, it's not entirely surprising that a certain degree of artifice might surround the Oscars. The academy insists that a contract signed by oscar winners mandates that they will not sell the trophies without first offering them to the academy for a nominal fee.the oscar statuette, one of the most recognizable images in the entertainment world, has no copyright protection, a federal judge has ruled. The walt disney co. filed suit in los angeles federal court thursday against the academy of motion picture arts and sciences, charging copyright infringement of its snow white character and unfair competition .

Cluster Name	Sentences Candidate	Date	Time
cluster0	``ben-hur'' in 1959 was the most-awarded film with 11, and walt disney was the most-awarded person with 32.	17-Feb-88	1039
cluster19	given the tremendous impact both an oscar nomination and an oscar victory have on a film's fortunes, it's not entirely surprising that a certain degree of artifice might surround the oscars.	23-Mar-89	1129
cluster2	the academy insists that a contract signed by oscar winners mandates that they will not sell the trophies without first offering them to the academy for a nominal fee.	24-Mar-89	733
cluster18	the oscar statuette, one of the most recognizable images in the entertainment world, has no copyright protection, a federal judge has ruled.	10-Nov-89	356
cluster4	in a lecture and question-and-answer session monday, lee shrugged off the academy's icy treatment of his latest movie, ``do the right thing,'' and blamed it on generational politics.	20-Feb-90	1033

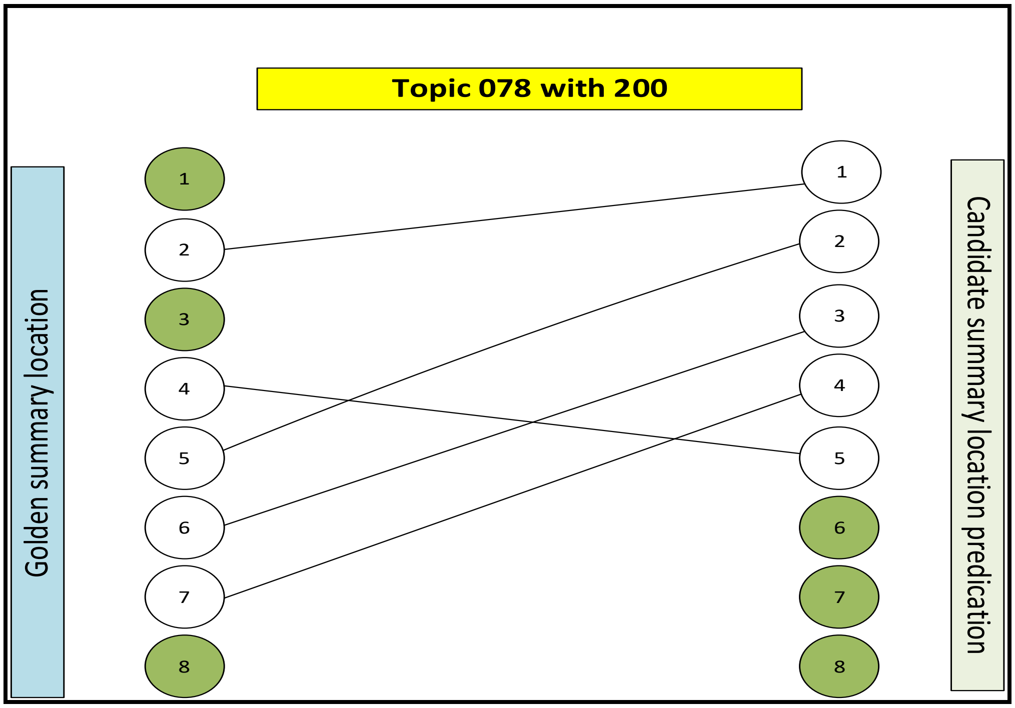

Figure 1 Compares Candidate Sentence Positions with Original Sentence Positions in Golden Summary Special To Topic 078 At 200 Words

Table 2 Compares Between the Correct Candidate Summary Sentences Prediction with The Golden Summary Sentences at Test 400 Words in Topic 078

the motion picture industry's most coveted award, oscar, was created 60 years ago and 1,816 of the statuettes have been produced so far. oscar, manufactured by the r.s.owens co., chicago, is made of britannia metal, copper plate, nickel plate and gold plate. last year, the oscarcast was seen by one billion people in 79 countries.``ben-hur'' in 1959 was the most-awarded film with 11, and walt disney was the most-awarded person with 32 .what actress has received the most oscars?katharine hepburn, 4. oscar's 60-year history is filled with examples of the film world's highest achievers being overlooked by the academy of motion picture arts and sciences. the academy appeared to make amends last year by presenting spielberg with the irving thalberg award for ``consistently high quality of production''. in a lecture and question-and- answer session monday, lee shrugged off the academy's icy treatment of his latest movie, ``do the right thing,'' and blamed it on generational politics. as it was, 10 different movies split 17 awards, the most winners for one year in this decade. the industry is still adjusting to its changing -- its older and more demanding audience -- and that means a wider variety of films. nominations for acting, directing, art direction and all specialized categories are handled by select branches and committees composed of academy members. all voting members of the academy can participate in the best picture nominations and may vote for the final awards in all categories. given the tremendous impact both an oscar nomination and an oscar victory have on a film’s fortunes, it's not entirely surprising that a certain degree of artifice might surround the oscars. the academy insists that a contract signed by oscar winners mandates that they will not sell the trophies without first offering them to the academy for a nominal fee.``but now that the word is out that these awards are worth $10,000 to $50,000, people are reluctant to sell them back to the academy for $10,'' willits said. the academy claimed that the star award, the trophy look-alike made by creative house promotions, violated copyright laws, diluted the academy's trademark and represented unfair competition. the court found otherwise. the walt disney co. filed suit in los angeles federal court thursday against the academy of motion picture arts and sciences, charging copyright infringement of its snow white character and unfair competition .

Cluster Name	Sentences Candidate	Date	Time
cluster0	the motion picture industry's most coveted award, oscar, was created 60 years ago and 1,816 of the statuettes have been produced so far.	17-Feb-88	1039
cluster0	``ben-hur'' in 1959 was the most-awarded film with 11, and walt disney was the most-awarded person with 32.	17-Feb-88	1039
cluster0	oscar, manufactured by the r.s.	17-Feb-88	1039
cluster9	owens co., chicago, is made of britannia metal, copper plate, nickel plate and gold plate.	17-Feb-88	1039
cluster6	the academy appeared to make amends last year by presenting spielberg with the irving thalberg award for ``consistently high quality of production.''	25-Mar-88	908
cluster11	last year, the oscarcast was seen by one billion people in 79 countries.	28-Mar-88	1345
cluster15	all voting members of the academy can participate in the best picture nominations and may vote for the final awards in all categories.	23-Mar-89	1129
cluster19	given the tremendous impact both an oscar nomination and an oscar victory have on a film's fortunes, it's not entirely surprising that a certain degree of artifice might surround the oscars.	23-Mar-89	1129
cluster10	nominations for acting, directing, art direction and all specialized categories are handled by select branches and committees composed of academy members.	23-Mar-89	1129
cluster2	the academy insists that a contract signed by oscar winners mandates that they will not sell the trophies without first offering them to the academy for a nominal fee.	24-Mar-89	0733
cluster12	``but now that the word is out that these awards are worth $10,000 to $50,000, people are reluctant to sell them back to the academy for $10,'' willits said.	24-Mar-89	0733
cluster16	the court found otherwise.	24-Mar-89	0733
cluster11	as it was, 10 different movies split 17 awards, the most winners for one year in this decade.	30-Mar-89	2400
cluster21	the industry is still adjusting to its changing -- its older and more demanding audience -- and that means a wider variety of films.	30-Mar-89	2400
cluster14	the academy claimed that the star award, the trophy look-alike made by creative house promotions, violated copyright	10-Nov-89	0356

	laws, diluted the academy's trademark and represented unfair competition.		
cluster4	in a lecture and question-and-answer session monday, lee shrugged off the academy's icy treatment of his latest movie, ``do the right thing,'' and blamed it on generational politics.	20-Feb-90	1033

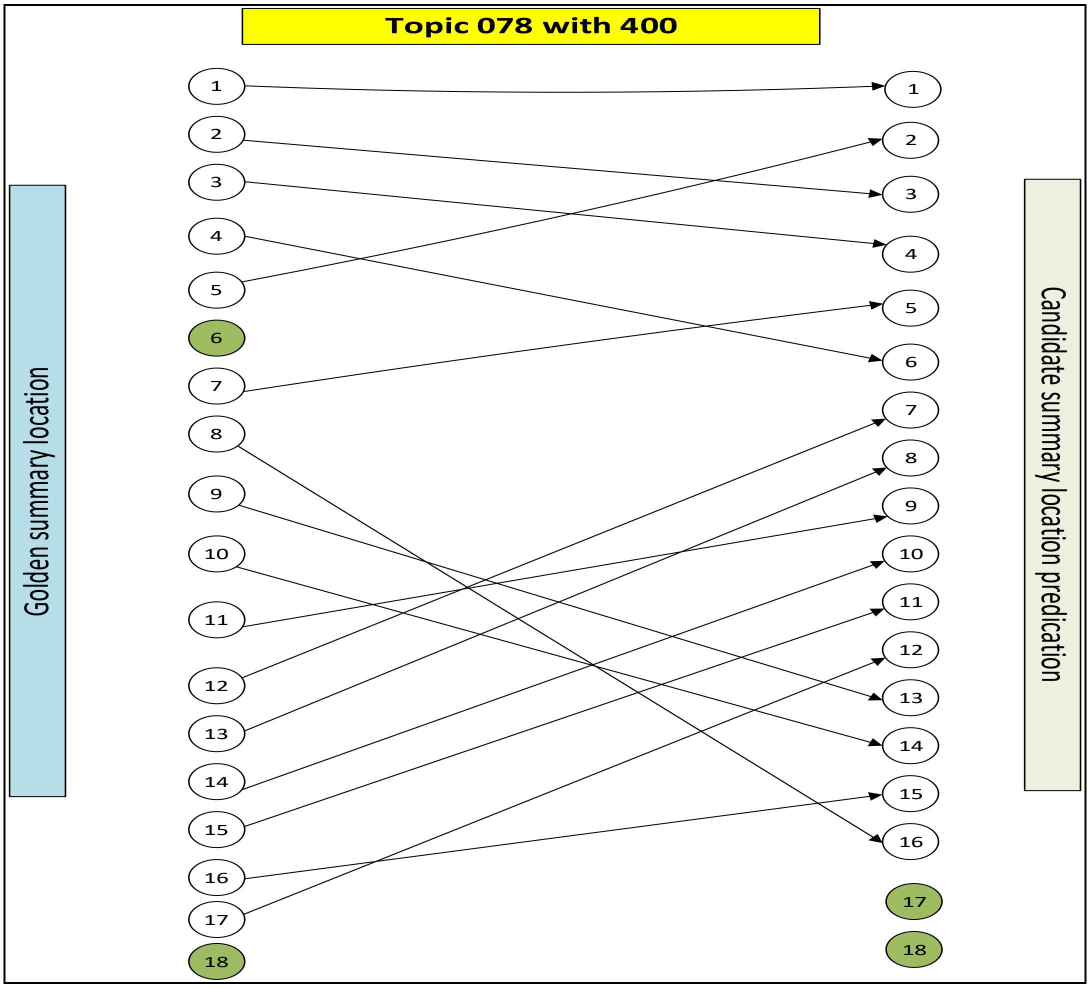

Figure 2 Compares Candidate Sentence Positions with Original Sentence Positions in Golden Summary Special to Topic 078 At 200 Words

Table 3 Compares Between the Correct Candidate Summary Sentences Prediction with The Golden Summary Sentences at Test 200 Words in Topic 087

calgary, alberta (ap). the best america could do was six medals, its worst winter games showing in 52 years. seoul, south korea (ap). the soviets took home 132 medals, including 55 gold, the most ever in a summer olympics without a major-power boycott. the united states finished third in medals. the summer olympics will be remembered for moments of glory like that enjoyed by

<table>
<tr><td colspan="4">u.s. diver greg louganis and the startling moment of disgrace when the gold was stripped from canadian sprinter ben johnson. olympic gold falls in advertising value --- medal winners lack appeal as endorsers. dave johnson of the united states clinched the decathlon gold medal in the final event at the world university games sunday .the late jesse owens, whose performance at the 1936 olympics put the lie to hitler's boasts of racial superiority, picked up a fifth gold medal wednesday for ``humanitarian contributions in the race of life''. shooter bakes sets record, wins gold. what the relays demonstrate more clearly than any other athletics event is that at the olympics the race has traditionally gone to affluent countries - or to those, rather, that choose to lavish resources on athletes.</td></tr>
<tr><td>Cluster Name</td><td>Sentences Candidate</td><td>Date</td><td>Time</td></tr>
<tr><td>cluster5</td><td>the best america could do was six medals, its worst winter games showing in 52 years.</td><td>28-Feb-88</td><td>2231</td></tr>
<tr><td>cluster5</td><td>the soviets took home 132 medals, including 55 gold, the most ever in a summer olympics without a major-power boycott.</td><td>2-Oct-88</td><td>1459</td></tr>
<tr><td>cluster7</td><td>the united states finished third in medals.</td><td>2-Oct-88</td><td>1459</td></tr>
<tr><td>cluster0</td><td>the summer olympics will be remembered for moments of glory like that enjoyed by u.s. diver greg louganis and the startling moment of disgrace when the gold was stripped from canadian sprinter ben johnson.</td><td>3-Oct-88</td><td>522</td></tr>
<tr><td>cluster22</td><td>dave johnson of the united states clinched the decathlon gold medal in the final event at the world university games sunday.</td><td>28-Aug-89</td><td>2400</td></tr>
<tr><td>cluster15</td><td>the late jesse owens, whose performance at the 1936 olympics put the lie to hitler's boasts of racial superiority, picked up a fifth gold medal wednesday for ``humanitarian contributions in the race of life.''</td><td>28-Mar-90</td><td>1820</td></tr>
<tr><td>cluster17</td><td>what the relays demonstrate more clearly than any other athletics event is that at the olympics the race has traditionally gone to affluent countries - or to those, rather, that choose to lavish resources on athletes.</td><td>8-Aug-92</td><td>2400</td></tr>
</table>

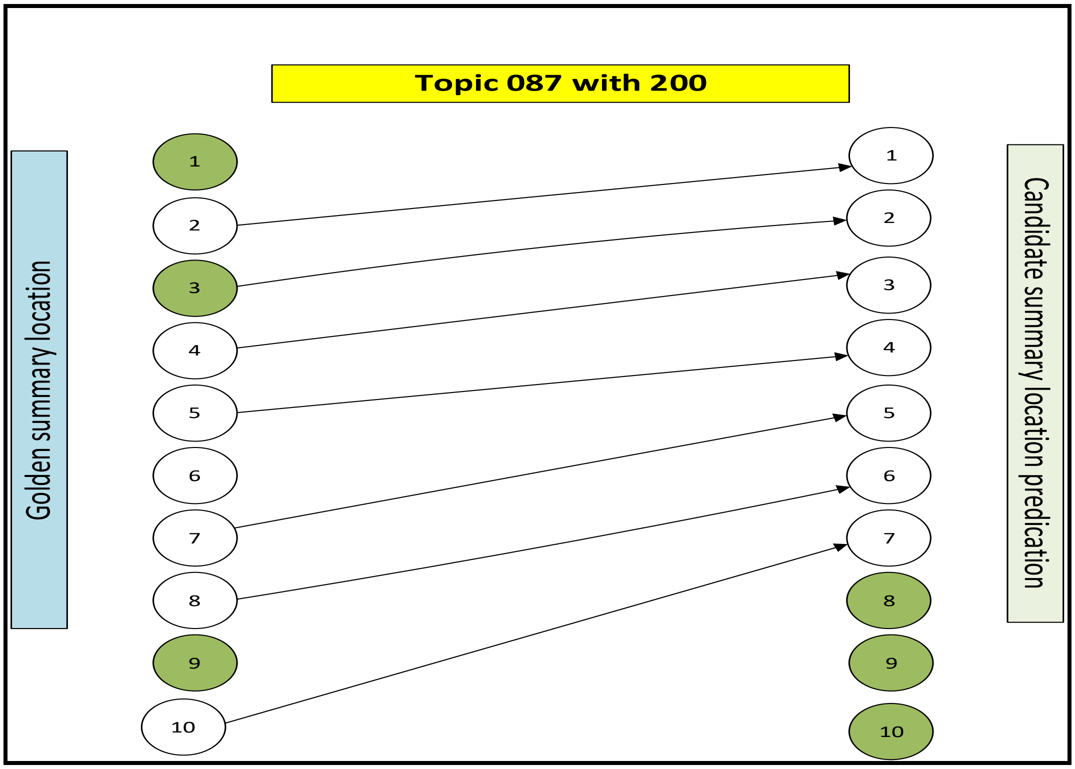

Figure 3 Compares Candidate Sentence Positions With Original Sentence Positions In Golden Summary Special To Topic 078 At 200 Words

Tabel 5 Compares Between the Correct Candidate Summary Sentences Prediction With The Golden Summary Sentences At Test 400 Words In Topic 087

<sum calgary, alberta (ap).debi thomas' dream of olympic gold turned into disappointment saturday as east germany's katarina witt won her second straight olympic championship and canadian elizabeth manley took home the silver before a crowd of cheering countrymen.the best america could do was six medals, its worst winter games showing in 52 years.seoul, south korea (ap).noted for her long, painted fingernails and racy, colorful bodysuits, griffith joyner of los angeles has a chance to match the 1984 quadruple gold medal performance of u.s. sprinter and long jumper carl lewis, and by fanny blankers-koen of the netherlands in 1948 .the soviets took home 132 medals, including 55 gold, the most ever in a summer olympics without a major-power boycott.the united states finished third in medals.the summer olympics will be remembered for moments of glory like that enjoyed by u.s. diver greg louganis and the startling moment of disgrace when the gold was stripped from canadian sprinter ben johnson.olympic gold falls in advertising value --- medal winners lack appeal as endorsers.now, marketers say, these medalists will be lucky to pull in $100,000 to $500,000 a year.dave johnson of the united states clinched the decathlon gold medal in the final event at the world university games sunday .the late jesse owens, whose performance at the 1936 olympics put the lie to hitler's boasts of racial superiority, picked up a fifth gold medal wednesday for ``humanitarian contributions in the race of life''.president bush presented the congressional gold medal to his widow, ruth owens, in a ceremony at the white house also attended by three daughters and teammates of the track legend.shooter bakes sets record, wins gold.bakes,

<table>
<tr><td colspan="4">27, set a world record in the kneeling portion of the three-position rifle competition friday en route to a gold medal at the 1990 world cup usa olympic-style shooting tournament at petersen's prado tiro ranges in chino .tonight in the big stadium in barcelona the athletics is dominated by the four relay finals.what the relays demonstrate more clearly than any other athletics event is that at the olympics the race has traditionally gone to affluent countries - or to those, rather, that choose to lavish resources on athletes.the cuban women's 100 metres relay teams of 1968 and 1972 are the only women from outside the developed world to have won relay medals.</td></tr>
<tr><td>Cluster Name</td><td>Sentences Candidate</td><td>Date</td><td>Time</td></tr>
<tr><td>cluster0</td><td>debi thomas' dream of olympic gold turned into disappointment saturday as east germany's katarina witt won her second straight olympic championship and canadian elizabeth manley took home the silver before a crowd of cheering countrymen.</td><td>28-Feb-88</td><td>0009</td></tr>
<tr><td>cluster5</td><td>the best america could do was six medals, its worst winter games showing in 52 years.</td><td>28-Feb-88</td><td>2231</td></tr>
<tr><td>cluster9</td><td>noted for her long, painted fingernails and racy, colorful bodysuits, griffith joyner of los angeles has a chance to match the 1984 quadruple gold medal performance of u.s. sprinter and long jumper carl lewis, and by fanny blankers-koen of the netherlands in 1948.</td><td>29-Sep-88</td><td>336</td></tr>
<tr><td>cluster5</td><td>the soviets took home 132 medals, including 55 gold, the most ever in a summer olympics without a major-power boycott.</td><td>2-Oct-88</td><td>1459</td></tr>
<tr><td>cluster7</td><td>the united states finished third in medals.</td><td>2-Oct-88</td><td>1459</td></tr>
<tr><td>cluster0</td><td>the summer olympics will be remembered for moments of glory like that enjoyed by u.s. diver greg louganis and the startling moment of disgrace when the gold was stripped from canadian sprinter ben johnson.</td><td>3-Oct-88</td><td>0522</td></tr>
<tr><td>cluster5</td><td>now, marketers say, these medalists will be lucky to pull in $100,000 to $500,000 a year.</td><td>4-Oct-88</td><td>2400</td></tr>
<tr><td>cluster22</td><td>dave johnson of the united states clinched the decathlon gold medal in the final event at the world university games sunday.</td><td>28-Aug-89</td><td>2400</td></tr>
<tr><td>cluster16</td><td>president bush presented the congressional gold medal to his widow, ruth owens, in a ceremony at the white house also attended by three daughters and teammates of the track legend.</td><td>28-Mar-90</td><td>1820</td></tr>
<tr><td>cluster15</td><td>the late jesse owens, whose performance at the 1936 olympics put the lie to hitler's boasts of racial superiority, picked up a fifth gold medal wednesday for ``humanitarian contributions in the race of life.''</td><td>28-Mar-90</td><td>1820</td></tr>
<tr><td>cluster22</td><td>bakes, 27, set a world record in the kneeling portion of the three-position rifle competition friday en route to a gold medal at the 1990 world cup usa olympic-style shooting tournament at petersen's prado tiro ranges in chino.</td><td>7-Apr-90</td><td>2400</td></tr>
<tr><td>cluster17</td><td>what the relays demonstrate more clearly than any other athletics event is that at the olympics the race has traditionally</td><td>8-Aug-92</td><td>2400</td></tr>
</table>

	gone to affluent countries - or to those, rather, that choose to lavish resources on athletes.		
cluster19	the cuban women's 100 metres relay teams of 1968 and 1972 are the only women from outside the developed world to have won relay medals.	8-Aug-92	2400
cluster20	tonight in the big stadium in barcelona the athletics is dominated by the four relay finals.	8-Aug-92	2400

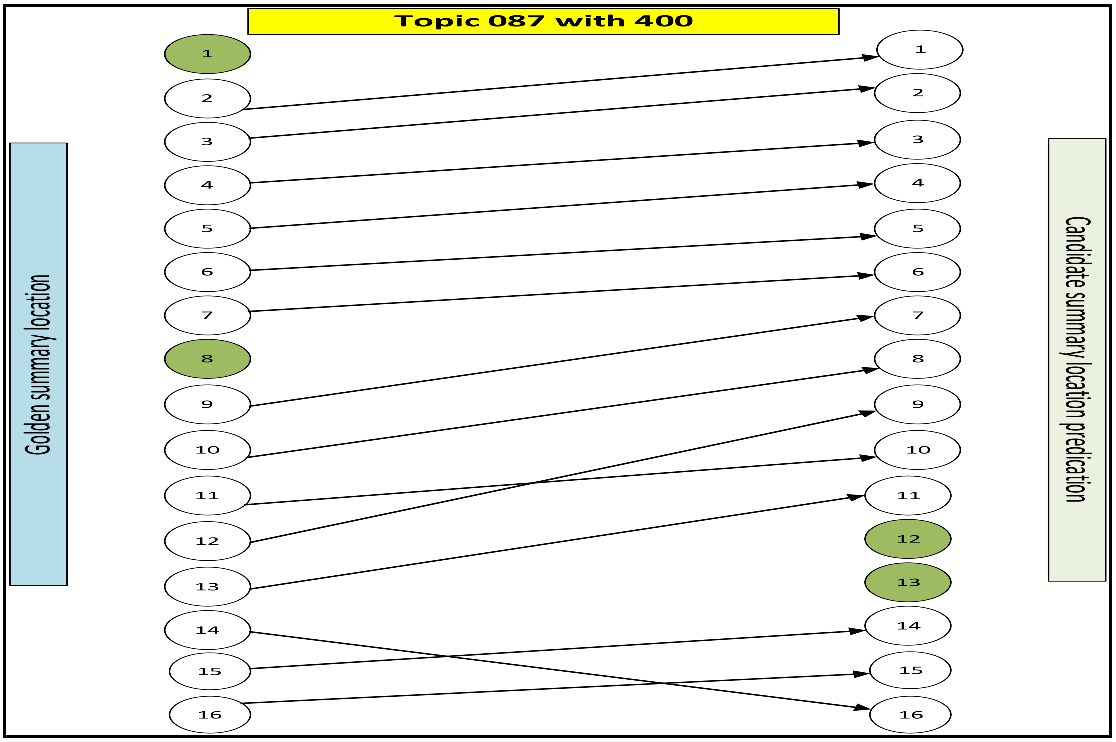

Figure 5 Compares Candidate Sentence Positions with Original Sentence Positions In Golden Summary Special To Topic 078 At 200 Words

Table 6 Compares Between the Correct Candidate Summary Sentences Prediction with The Golden Summary Sentences at Test 200 Words in Topic 090

handling of the controversy highlighted all the inconsistencies for which mrs. aquino has been criticized: lack of clear priorities tactical blundering and the strong influence of her relatives over national policy.president corazon aquino said today she will not allow ferdinand marcos to come home to die and her government will not permit the ousted president to be buried on philippine soil.mrs. aquino said the philippines' debt was projected to increase to $29 billion by year's end but that much was money that ``was in fact stolen by the previous government'' of ousted dictator ferdinand marcos.as a military coup attempt unfolded in their homeland 7,000 miles away, supporters of the rebellious soldiers engaged in a shouting match thursday night with supporters of president corazon aquino at the philippine consulate in los angeles .president corazon aquino today reversed a long-held policy and said she was willing to negotiate cease-

fires with communist guerrillas and former soldiers who have staged coup attempts. the quick collapse of a bizarre military revolt on the philippines' second-largest island leaves president corazon aquino's beleaguered government with a muddied victory at best and clear danger signs ahead .

Cluster Name	Sentences Candidate	Date	Time
cluster0	president corazon aquino said today she will not allow ferdinand marcos to come home to die and her government will not permit the ousted president to be buried on philippine soil.	19-May-89	0707
cluster0	mrs. aquino said the philippines' debt was projected to increase to $29 billion by year's end but that much was money that ``was in fact stolen by the previous government'' of ousted dictator ferdinand marcos.	10-Jul-89	2059
cluster11	handling of the controversy highlighted all the inconsistencies for which mrsaquino has been criticized: lack of clear priorities tactical blundering and the strong influence of her relatives over national policy.	8-Apr-90	2117
cluster12	president corazon aquino today reversed a long-held policy and said she was willing to negotiate cease-fires with communist guerrillas and former soldiers who have staged coup attempts.	29-Aug-90	0838

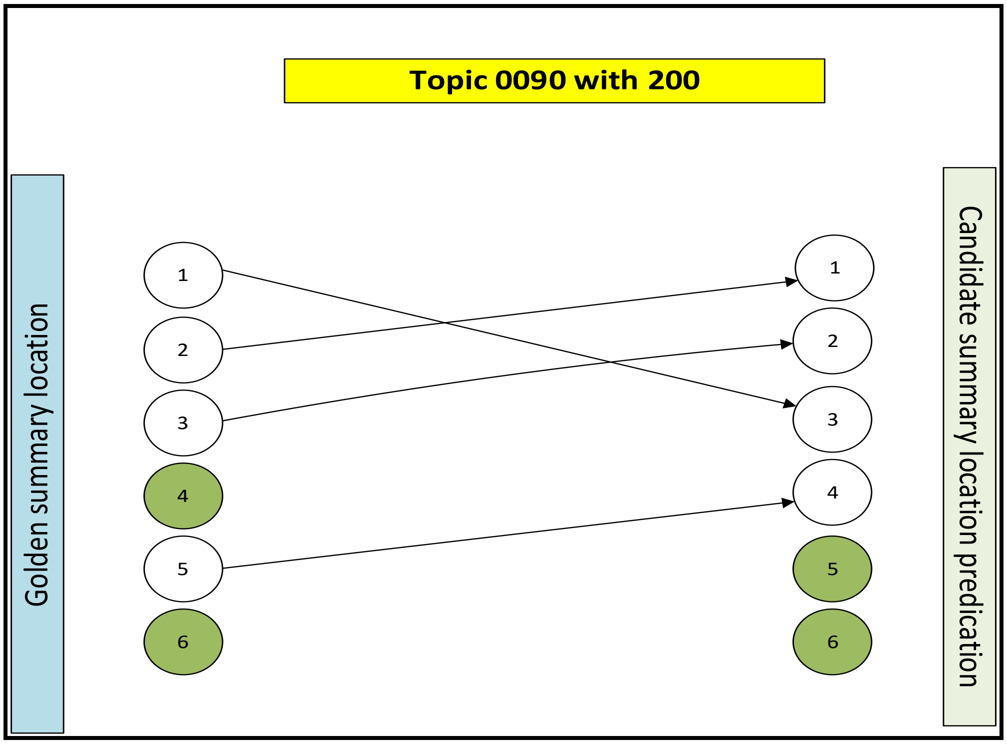

Figure 6 Compares Candidate Sentence Positions With Original Sentence Positions In Golden Summary Special To Topic 090 At 200 Words

Table 7 Compares Between the Correct Candidate Summary Sentences Prediction With The Golden Summary Sentences At Test 400 Words In Topic 090

handling of the controversy highlighted all the inconsistencies for which mrs. aquino has been criticized: lack of clear priorities tactical blundering and the strong influence of her relatives over national policy.president corazon aquino has agreed to allow ferdinand marcos to return to the philippines so that the government can try him on charges of stealing billions of dollars, newspapers said saturday.president corazon aquino said friday that for security reasons she would not permit deposed president ferdinand e. marcos to be buried in the philippines.government officials have accused marcos supporters of taking part in at least two attempted coups against the aquino government .aquino evidently is concerned that even after death, the return of marcos' remains could spark violence or galvanize opposition to her government .the 71-year-old marcos was in critical condition friday in a honolulu hospital after emergency kidney surgery.mrs.aquino, on the first stop of her weeklong three-nation european tour, met with west german president richard von weizsaecker, foreign minister hans-dietrich genscher and other officials.mrs. aquino said the philippines' debt was projected to increase to $29 billion by year's end but that much was money that ``was in fact stolen by the previous government'' of ousted dictator ferdinand marcos.philippine vice president and opposition leader salvador h. laurel declined saturday night to condemn the armed insurrection in his country and said the rebels "have the right" to try to seize power .as a military coup attempt unfolded in their homeland 7,000 miles away, supporters of the rebellious soldiers engaged in a shouting match thursday night with supporters of president corazon aquino at the philippine consulate in los angeles . the resignation of the land reform secretary raises doubts about president corazon aquino's commitment to agrarian reform and dramatizes conflicts in a government criticized for lack of vision.president corazon aquino today reversed a long-held policy and said she was willing to negotiate cease-fires with communist guerrillas and former soldiers who have staged coup attempts.the quick collapse of a bizarre military revolt on the philippines' second-largest island leaves president corazon aquino's beleaguered government with a muddied victory at best and clear danger signs ahead .ironically, noble made his move on the day a presidential fact-finding board issued a long-awaited 609-page report on the causes of last december's coup attempt in manila, which left more than 100 people dead .

Cluster Name	Sentences Candidate	Date	Time
cluster0	president corazon aquino has agreed to allow ferdinand marcos to return to the philippines so that the government can try him on charges of stealing billions of dollars, newspapers said saturday.	25-Jun-88	1555
cluster3	the 71-year-old marcos was in critical condition friday in a honolulu hospital after emergency kidney surgery.	19-May-89	1438
cluster0	president corazon aquino said friday that for security reasons she would not permit deposed president ferdinand e. marcos to be buried in the philippines.	20-May-89	2400
cluster0	aquino evidently is concerned that even after death, the return of marcos' remains could spark violence or galvanize opposition to her government.	20-May-89	2400
cluster0	mrs. aquino said the philippines' debt was projected to increase to $29 billion by year's end but that much was money that ``was in fact stolen by the previous government'' of ousted dictator ferdinand marcos.	10-Jul-89	2059
cluster9	mrs aquino, on the first stop of her weeklong three-nation european tour, met with west german president richard von weizsaecker, foreign minister hans-dietrich genscher and other officials.	10-Jul-89	2059
cluster19	philippine vice president and opposition leader salvador h. laurel declined saturday night to condemn the armed insurrection in his country and said the rebels "have the right" to try to seize power.	3-Dec-89	2400

cluster10	the resignation of the land reform secretary raises doubts about president corazon aquino's commitment to agrarian reform and dramatizes conflicts in a government criticized for lack of vision.	8-Apr-90	2117
cluster12	president corazon aquino today reversed a long-held policy and said she was willing to negotiate cease-fires with communist guerrillas and former soldiers who have staged coup attempts.	29-Aug-90	0838

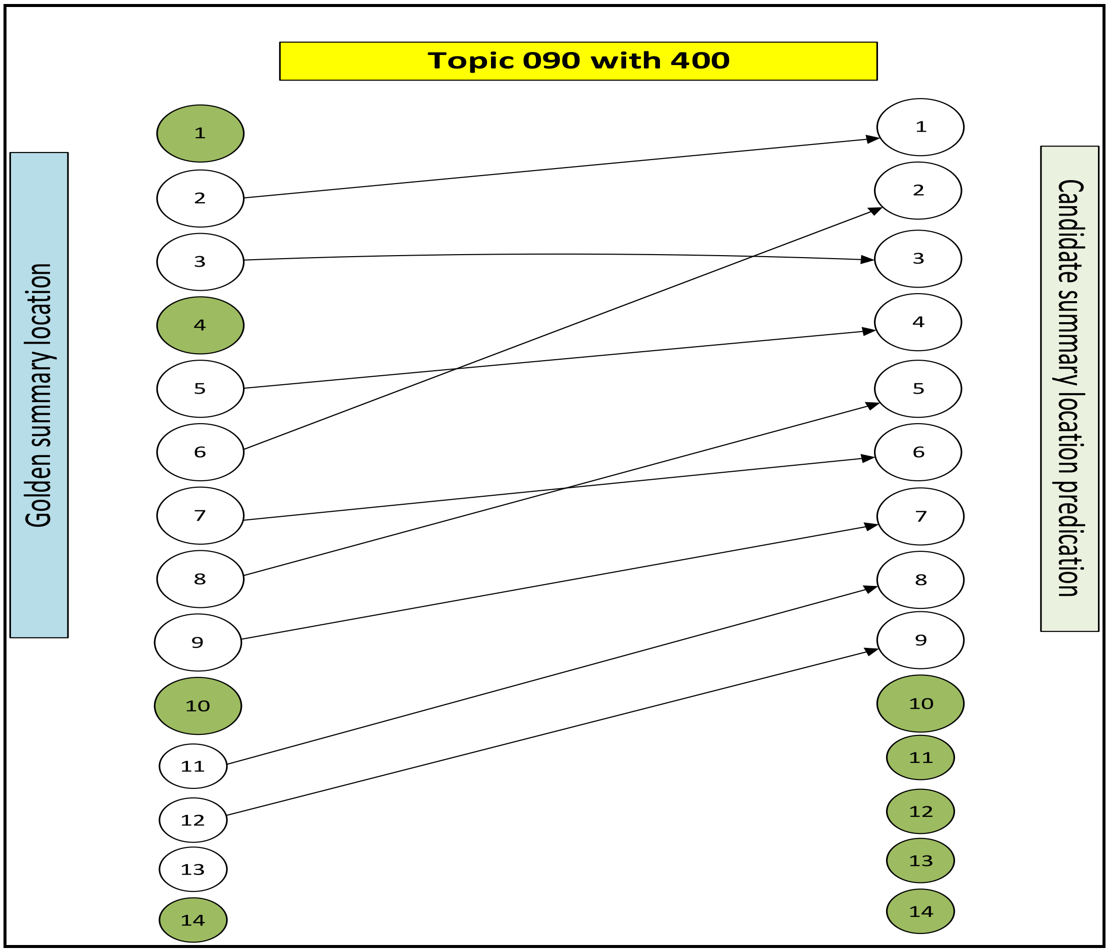

Figure 7 Shows Candidate Sentence Positions Compared With Original Sentence Positions In Golden Summary Special To Topic 090 At 400 Words

Table 8 Compare Between The Correct Candidate Summary Sentences Prediction With The Golden Summary Sentences At Test 200 Words In Topic 105

ethnic violence killed 148 people and wounded 885 this month in the central asian republic of kirghizia, the official news agency tass said wednesday.besides unrest this month in kirghizia and uzbekistan, ethnic violence also has been reported in recent years in georgia, armenia, azerbaijan, kazakhstan and tadzhikistan .tajikistan was hit most hard.the civil war in the republic claimed the lives of more than 100,000 people, shabdolov said.he said in moscow on tuesday that ministry troops had to be used in 13 regions of the soviet union because of such outbreaks, tass said.soviet president mikhail s. gorbachev

<table>
<tr><td colspan="4">warned the nation saturday night that the country's growing ethnic strife is endangering its unity as a state as well as threatening his program of political and economic reforms .gorbachev acknowledged that ethnic relations are "wound in a tight knot of contradictions" because of previous soviet policies, particularly the effort to pull the country's 100-plus ethnic groups into a new "soviet nationality" with a unified culture .the international community is starting to look beyond the conflict in the former yugoslavia, and is becoming increasingly aware of the true magnitude and seriousness of ethnic and political conflicts in the former soviet union.</td></tr>
<tr><td>Cluster Name</td><td>Sentences Candidate</td><td>Date</td><td>Time</td></tr>
<tr><td>cluster0</td><td>soviet president mikhail s. gorbachev warned the nation saturday night that the country's growing ethnic strife is endangering its unity as a state as well as threatening his program of political and economic reforms.</td><td>2-Jul-89</td><td>2400</td></tr>
<tr><td>cluster9</td><td>ethnic violence killed 148 people and wounded 885 this month in the central asian republic of kirghizia, the official news agency tass said wednesday.</td><td>13-Jun-90</td><td>2337</td></tr>
<tr><td>cluster2</td><td>he said in moscow on tuesday that ministry troops had to be used in 13 regions of the soviet union because of such outbreaks, tass said.</td><td>13-Jun-90</td><td>2337</td></tr>
<tr><td>cluster1</td><td>besides unrest this month in kirghizia and uzbekistan, ethnic violence also has been reported in recent years in georgia, armenia, azerbaijan, kazakhstan and tadzhikistan.</td><td>13-Jun-90</td><td>2337</td></tr>
<tr><td>cluster11</td><td>the international community is starting to look beyond the conflict in the former yugoslavia, and is becoming increasingly aware of the true magnitude and seriousness of ethnic and political conflicts in the former soviet union.</td><td>14-Dec-93</td><td>2400</td></tr>
<tr><td>cluster5</td><td>the civil war in the republic claimed the lives of more than 100,000 people, shabdolov said.</td><td>16-May-94</td><td>2400</td></tr>
<tr><td>cluster3</td><td>tajikistan was hit most hard.</td><td>16-May-94</td><td>2400</td></tr>
</table>

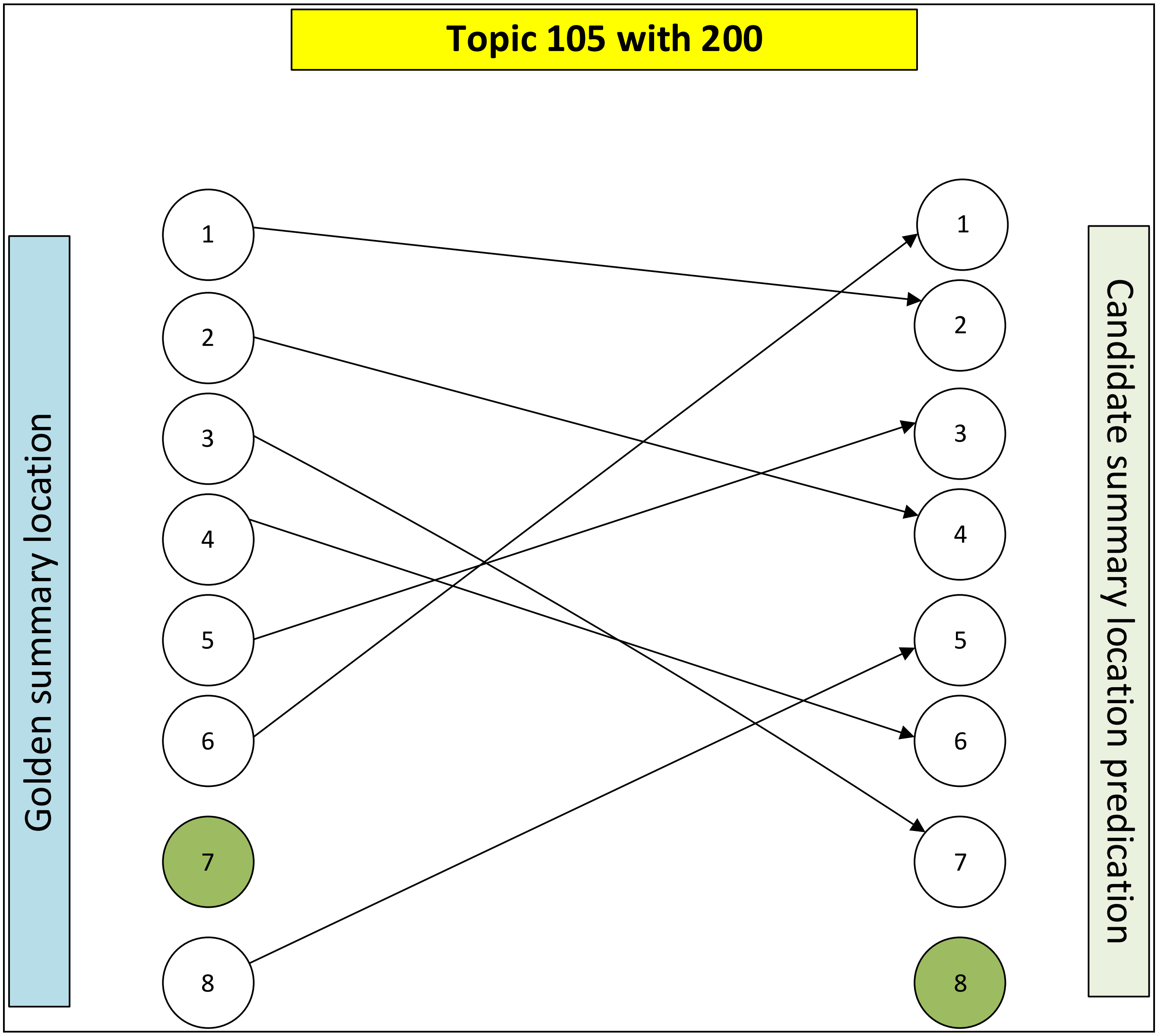

Figure 8 Displays Candidate Sentence Positions Compared With Original Sentence Positions In Golden Summary Special To Topic 105 At 200 Words

Table 9 compare Between The Correct Candidate Summary Sentences Prediction With The Golden Summary Sentences At Test 400 Words In Topic 105

state-run media also indicated friday that a crackdown had begun on people in armenia who continue to agitate for annexation of a mostly armenian enclave in the neighboring republic of azerbaijan.the government has decided to give ethnic germans and crimean tatars who were expelled from their homelands by josef stalin their own autonomous regions but it will take years, tass reported wednesday.it was also a potentially dangerous precedent: the soviet union is a patchwork of well over 100 ethnic groups, and many of them have outstanding territorial claims. ethnic violence killed 148 people and wounded 885 this month in the central asian republic of kirghizia, the official news agency tass said wednesday.he said in moscow on tuesday that ministry troops had to be used in 13 regions of the soviet

union because of such outbreaks, tass said.moldavia, acting to defuse one of the soviet union's most explosive ethnic conflicts, agreed sunday to comply with a kremlin order to reconsider its law making moldavian the republic's official language.tajikistan was hit most hard.the civil war in the republic claimed the lives of more than 100,000 people, shabdolov said.soviet president mikhail s. gorbachev warned the nation saturday night that the country's growing ethnic strife is endangering its unity as a state as well as threatening his program of political and economic reforms .soviet president mikhail s. gorbachev warned the nation saturday night that the country's growing ethnic strife is endangering its unity as a state as well as threatening his program of political and economic reforms .with serious ethnic clashes in soviet central asia over the past month, with troops needed to maintain the peace in the southern republics of armenia, azerbaijan and georgia and with nationalism growing stronger in the baltic republics of estonia, latvia and lithuania, gorbachev is struggling to hold the soviet union together as a state while he pushes broad reforms that he believes will resolve most of the grievances .gorbachev acknowledged that ethnic relations are "wound in a tight knot of contradictions" because of previous soviet policies, particularly the effort to pull the country's 100-plus ethnic groups into a new "soviet nationality" with a unified culture .the international community is starting to look beyond the conflict in the former yugoslavia, and is becoming increasingly aware of the true magnitude and seriousness of ethnic and political conflicts in the former soviet union.

Cluster Name	Sentences Candidate	Date	Time
cluster1	state-run media also indicated friday that a crackdown had begun on people in armenia who continue to agitate for annexation of a mostly armenian enclave in the neighboring republic of azerbaijan.	29-Jul-88	1643
cluster0	soviet president mikhail s. gorbachev warned the nation saturday night that the country's growing ethnic strife is endangering its unity as a state as well as threatening his program of political and economic reforms.	2-Jul-89	2400
cluster4	the government has decided to give ethnic germans and crimean tatars who were expelled from their homelands by josef stalin their own autonomous regions but it will take years, tass reported wednesday.	13-Dec-89	1711
cluster2	it was also a potentially dangerous precedent: the soviet union is a patchwork of well over 100 ethnic groups, and many of them have outstanding territorial claims.	13-Dec-89	1711
cluster9	ethnic violence killed 148 people and wounded 885 this month in the central asian republic of kirghizia, the official news agency tass said wednesday.	13-Jun-90	2337
cluster2	he said in moscow on tuesday that ministry troops had to be used in 13 regions of the soviet union because of such outbreaks, tass said.	13-Jun-90	2337
cluster0	moldavia, acting to defuse one of the soviet union's most explosive ethnic conflicts, agreed sunday to comply with a kremlin order to reconsider its law making moldavian the republic's official language.	30-Dec-90	1601
cluster11	the international community is starting to look beyond the conflict in the former yugoslavia, and is becoming increasingly aware of the true magnitude and seriousness of ethnic and political conflicts in the former soviet union.	14-Dec-93	2400
cluster5	the civil war in the republic claimed the lives of more than 100,000 people, shabdolov said.	16-May-94	2400
cluster3	tajikistan was hit most hard.	16-May-94	2400

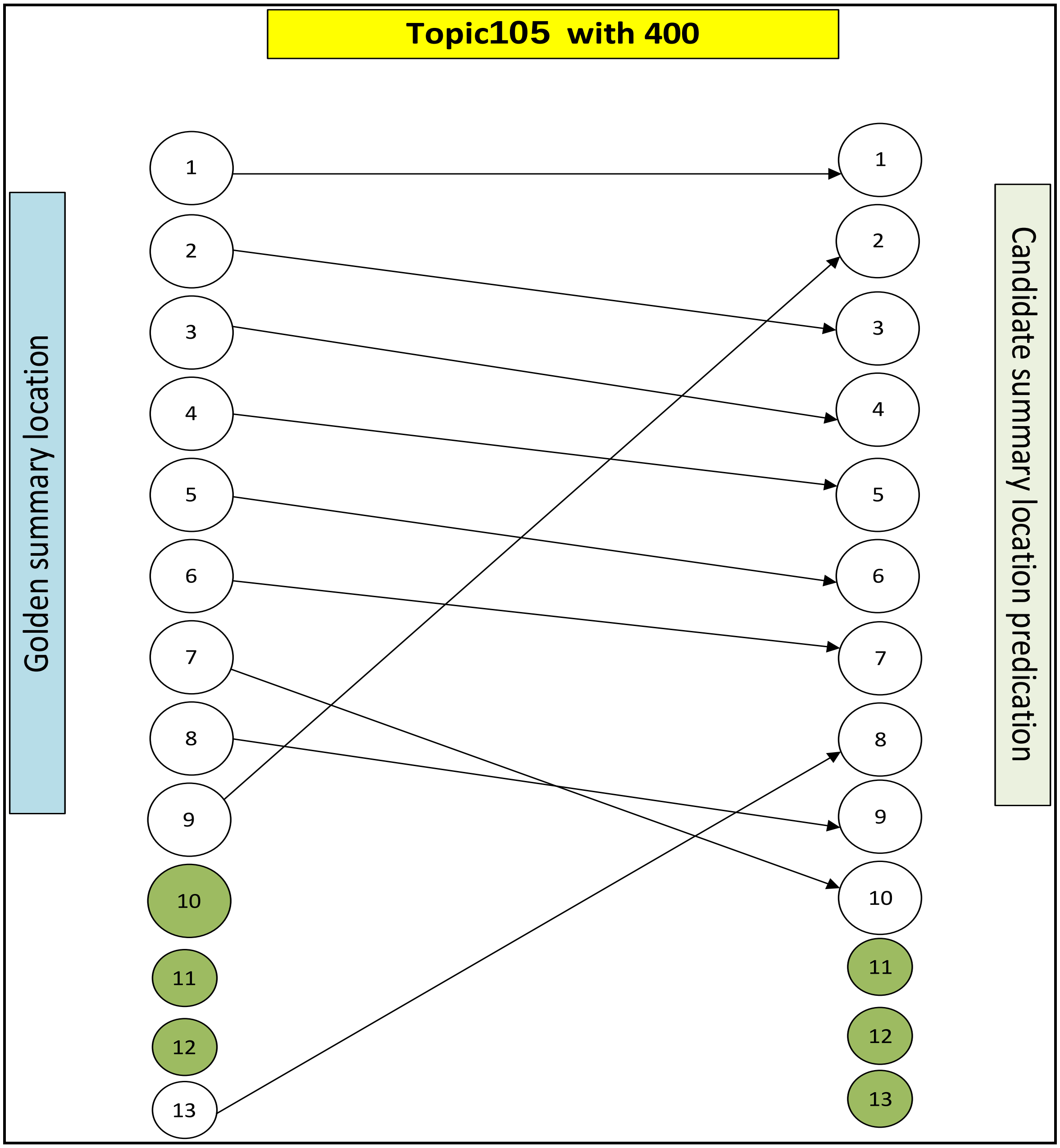

Figure 9 Displays Candidate Sentence Positions Compared With Original Sentence Positions In Golden Summary Special To Topic 105 At 200 Words

Table 10 Compare Between The Correct Candidate Summary Sentences Prediction With The Golden Summary Sentences At Test 200 Words In Topic 119

gildenhorn, according to senate sources, is one of several bush ambassadorial nominees who have been rated``unqualified'' by the american academy of diplomacy, an organization composed of former high ranking diplomats, including all living former secretaries of state.other ambassadorial nominees rated as unqualified by the american academy of diplomacy include florida real estate developer joseph zappala, nominated as ambassador to spain; melvin l. sembler, a real estate executive in florida named to be ambassador to australia; and della m. newman, a seattle, wash., real estate broker named to be ambassador to new zealand.the bush administration is expected to name career diplomat harry

<table>
<tr><td colspan="4">shlaudeman as the first u.s. ambassador to nicaragua in almost two years, a u.s. official says.the choice is james r. lilley, a career intelligence officer who served as cia station chief in china when bush headed the u.s. liaison office there in 1974-75 .president bush has decided to nominate john d. negroponte, a veteran diplomat who helped direct u.s. aid to nicaraguan rebels, to the key position of ambassador to mexico, administration officials said thursday .washington -- president bush plans to name united nations ambassador thomas pickering as u.s. envoy to india, and appoint the current head of the foreign service to succeed him.;</td></tr>
<tr><td>Cluster Name</td><td>Sentences Candidate</td><td>Date</td><td>Time</td></tr>
<tr><td>cluster5</td><td>president bush has decided to nominate john d. negroponte, a veteran diplomat who helped direct u.s. aid to nicaraguan rebels, to the key position of ambassador to mexico, administration officials said thursday.</td><td>3-Feb-89</td><td>2400</td></tr>
<tr><td>cluster1</td><td>gildenhorn, according to senate sources, is one of several bush ambassadorial nominees who have been rated ``unqualified'' by the american academy of diplomacy, an organization composed of former high ranking diplomats, including all living former secretaries of state.</td><td>15-Jul-89</td><td>1301</td></tr>
<tr><td>cluster3</td><td>other ambassadorial nominees rated as unqualified by the american academy of diplomacy include florida real estate developer joseph zappala, nominated as ambassador to spain melvin l. sembler, a real estate executive in florida named to be ambassador to Australia and della m. newman, a seattle, wash., real estate broker named to be ambassador to new zealand.</td><td>15-Jul-89</td><td>1301</td></tr>
<tr><td>cluster5</td><td>the bush administration is expected to name career diplomat harry shlaudeman as the first u.s. ambassador to nicaragua in almost two years, a u.s. official says.</td><td>10-May-90</td><td>303</td></tr>
<tr><td>cluster5</td><td>washington -- president bush plans to name united nations ambassador thomas pickering as u.s. envoy to india, and appoint the current head of the foreign service to succeed him.</td><td>3-Feb-92</td><td>2400</td></tr>
</table>

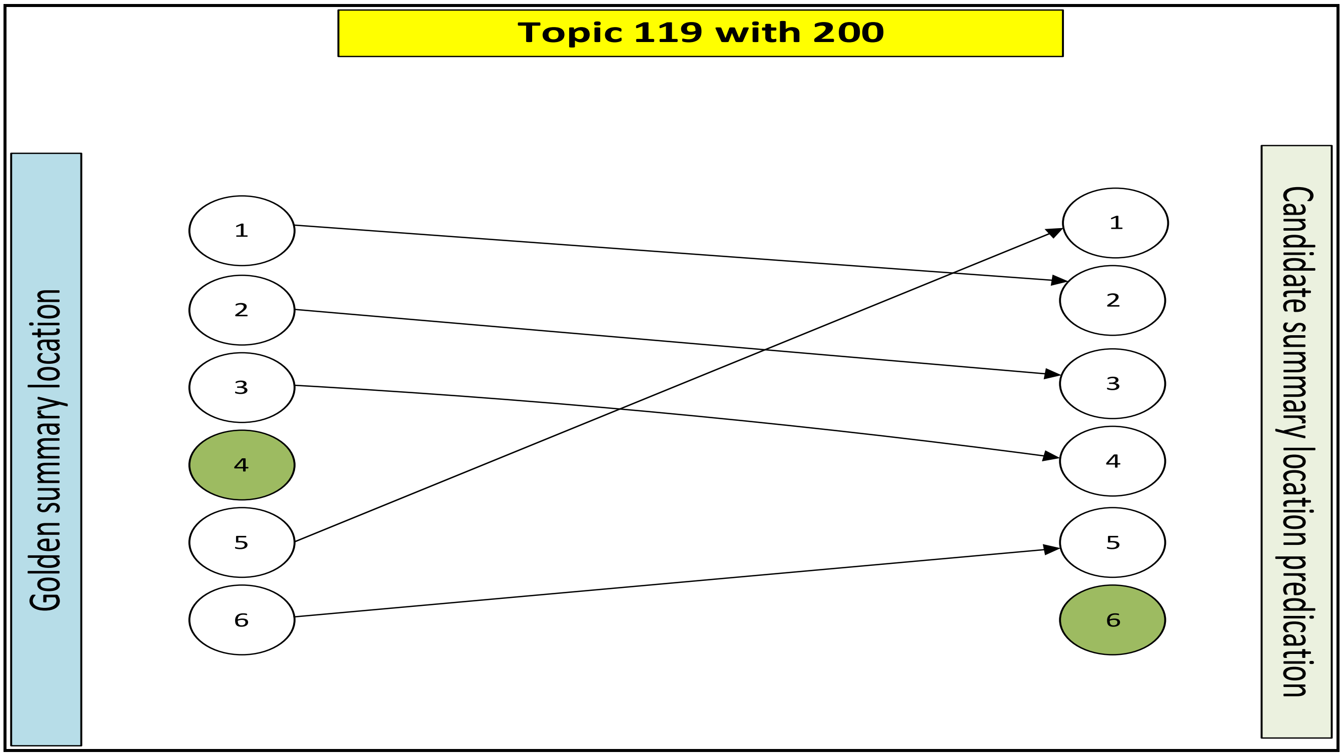

Figure 10 Displays Candidate Sentence Positions Compared With Original Sentence Positions In Golden Summary Special To Topic 119 At 200 Words

Table 11 Compare Between The Correct Candidate Summary Sentences Prediction With The Golden Summary Sentences At Test 400 Words In Topic 119

The Senate Foreign Relations Committee today endorsed the nomination of Michigan Republican activist Peter F. Secchia as ambassador to Italy, but the panel's chairman said he feared Secchia might ``embarrass the United States'' because of his ``penchant for making gross remarks''.
Gildenhorn, according to Senate sources, is one of several Bush ambassadorial nominees who have been rated ``unqualified'' by the American Academy of Diplomacy, an organization composed of former high ranking diplomats, including all living former secretaries of state.
Other ambassadorial nominees rated as unqualified by the American Academy of Diplomacy include Florida real estate developer Joseph Zappala, nominated as ambassador to Spain; Melvin L. Sembler, a real estate executive in Florida named to be ambassador to Australia; and Della M. Newman, a Seattle, Wash., real estate broker named to be ambassador to New Zealand.
Ronald I. Spiers, a veteran U.S.diplomat, will be the new undersecretary-general for General Assembly affairs, the top-ranking American post at the world body, U.N. officials said Tuesday.
Hinton, currently ambassador to Costa Rica, replaces Ambassador Arthur H. Davis, who was recalled by Bush in protest of what the administration considered the stealing of the Panamanian elections last May by Gen. Manuel Antonio Noriega.
The Bush administration is expected to name career diplomat Harry Shlaudeman as the first U.S. ambassador to Nicaragua in almost two years, a U.S. official says.
The administration had planned to send another career diplomat, Melissa Wells, to Managua but had a change of heart after concluding she might face a prolonged Senate confirmation fight because she has been a target of conservatives.

The choice is James R. Lilley, a career intelligence officer who served as CIA station chief in China when Bush headed the U.S. liaison office there in 1974-75 .

President Bush has decided to nominate John D. Negroponte, a veteran diplomat who helped direct U.S. aid to Nicaraguan rebels, to the key position of ambassador to Mexico, Administration officials said Thursday .

It also approved the nominations of Thomas Melady to be ambassador to the Vatican, William Howard Taft IV to be the permanent U.S. representative on the Council of the North Atlantic Treaty Organization, Keith L. Brown to be ambassador to Denmark, and Joseph Gildenhorn, a Washington lawyer and real estate developer, to be ambassador to Switzerland .

WASHINGTON -- President Bush plans to name United Nations Ambassador Thomas Pickering as U.S. envoy to India, and appoint the current head of the foreign service to succeed him.

cluster3	it also approved the nominations of thomas melady to be ambassador to the vatican, william howard taft iv to be the permanent u.s. representative on the council of the north atlantic treaty organization, keith l. brown to be ambassador to denmark, and joseph gildenhorn, a washington lawyer and real estate developer, to be ambassador to switzerland.	11-Sep-86	2400
cluster5	president bush has decided to nominate john d. negroponte, a veteran diplomat who helped direct u.s. aid to nicaraguan rebels, to the key position of ambassador to mexico, administration officials said thursday.	3-Feb-89	2400
cluster0	the senate foreign relations committee today endorsed the nomination of michigan republican activist peter f. secchia as ambassador to italy, but the panel's chairman said he feared secchia might ``embarrass the united states'' because of his ``penchant for making gross remarks.''	17-May-89	1321
cluster1	gildenhorn, according to senate sources, is one of several bush ambassadorial nominees who have been rated ``unqualified'' by the american academy of diplomacy, an organization composed of former high ranking diplomats, including all living former secretaries of state.	15-Jul-89	1301
cluster3	other ambassadorial nominees rated as unqualified by the american academy of diplomacy include florida real estate developer joseph zappala, nominated as ambassador to spain melvin l. sembler, a real estate executive in florida named to be ambassador to Australia and della m. newman, a seattle, wash., real estate broker named to be ambassador to new zealand.	15-Jul-89	1301
cluster5	ronald i. spiers, a veteran u.s. diplomat, will be the new undersecretary-general for general assembly affairs, the top-ranking american post at the world body, u.n. officials said tuesday.	1-Aug-89	2053
cluster6	hinton, currently ambassador to costa rica, replaces ambassador arthur h. davis, who was recalled by bush in protest of what the administration considered the stealing of the panamanian elections last may by gen. manuel antonio noriega.	2-Jan-90	1736
cluster5	the bush administration is expected to name career diplomat harry shlaudeman as the first u.s. ambassador to nicaragua in almost two years, a u.s. official says.	10-May-90	303
cluster11	the administration had planned to send another career diplomat, melissa wells, to managua but had a change of heart after concluding she might face a prolonged senate confirmation fight because she has been a target of conservatives.	10-May-90	303
cluster5	washington -- president bush plans to name united nations ambassador thomas pickering as u.s. envoy to india, and appoint the current head of the foreign service to succeed him.	3-Feb-92	2400

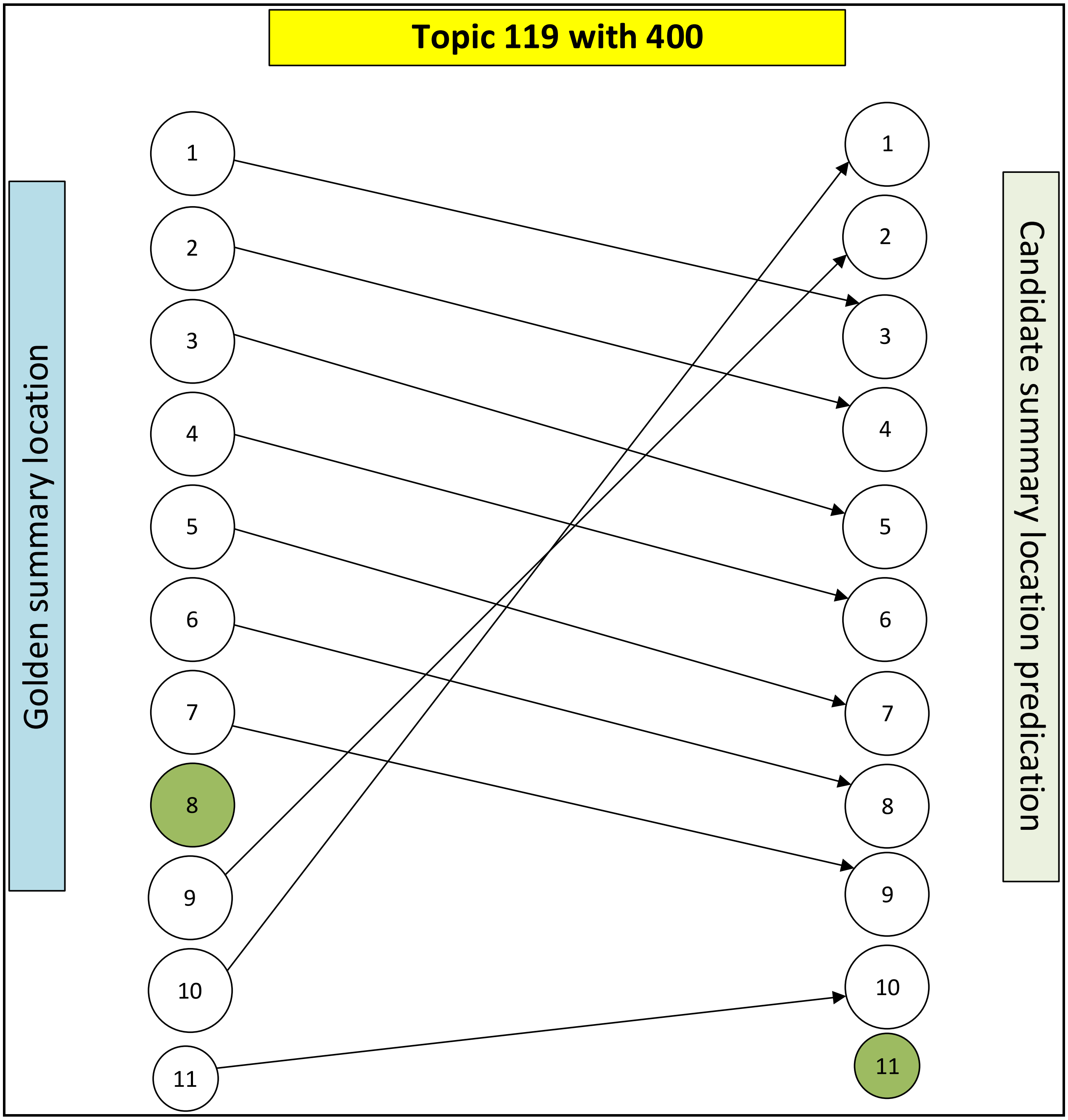

Figure 11 Displays Candidate Sentence Positions Compared with Original Sentence Positions In Golden Summary Special To Topic 119 At 400 Words

المستخلص

النمو المتزايد في حجم البيانات الرقمية للوثائق أدى إلى صعوبة الوصول إلى المعلومات المهمة. الحل هو استخدام أنظمة تلخيص آلية تهدف إلى استخراج المعلومات المهمة في وقت قصير. عادة ما يكون عمل هذه الأنظمة هو استخراج ملخص واحد من مستند واحد أو مستندات متعددة.

تقدم هذه الأطروحة عدة اتجاهات. الأول هو اقتراح خوارزمية جديدة صاغت اسم خوارزمية تطوير المجموعات (DCA) لجمع البيانات غير المسماة ووضعها في مجموعات مناسبة. والثاني هو إنشاء سلسلة معجمية بناءً على جمل دلالية متشابهة أو عدد مشابه من الكلمات بين الجمل التي صاغ اسمها معجم سلسلة الجمل (LCS) ، والتي تختلف عن سلسلة الكلمات المعجمية التقليدية (LCW) التي تعمل على أساس الكلمات. والثالث اقتراح مجموعة من الميزات لاستخراج جمل مهمة وسهلة الفهم. الرابع هو بناء شبكة عصبية متعددة الطبقات (BMPNN) للعثور على درجة الجملة. الخامس هو استخدام طريقة عشوائية (ROS) ودورها الفعال في إعادة موازنة البيانات أثناء عملية التدريب في BMPNN. أخيرًا ، يتم حل مشكلة إعادة ترتيب الجمل في ملخص المرشح وفقًا لأهمية الجملة بالاعتماد على التاريخ بالإضافة إلى ثلاثة شروط تؤخذ في الاعتبار لضمان دقة عملية إعادة الترتيب. استخدم هذا العمل مجموعتين من البيانات ذات الأهمية في المقالات الإخبارية. مجموعة البيانات الأولى هي مؤتمر فهم المستندات (DUC 2002) ، وقد تم إنشاء مجموعة البيانات الثانية يدويًا من الوثائق الإخبارية التي تم جمعها من قبلنا للتجارب.

أظهرت النتائج أن أداء خوارزمية DCA المقترحة قد تفوق عمومًا على خوارزمية التجميع الهرمي بعدد المجموعات المتولدة ، وأيضًا على خوارزمية K-mean من خلال نتائج التقييم الناتجة من مقياس Davies Bouldin Index (DBI). فيما يتعلق بتقييم ملخص المرشح. استخدمت هذه الرسالة ثلاثة مقاييس لمقاييس عائلة Recall-Oriented Understudy for Gisting Evaluation (Rouge) ، و Rouge-1 ، و Rouge-2 ، و Rouge-L لتقييم ملخص المرشح. أظهرت نتائج تقييم مقاييس Rouge-1 و Rouge-2 و Rouge-L أن ملخص المرشح قريب جدًا من الملخص الذهبي من حيث مطابقة الجمل ، وقد حقق نتائج واعدة. متوسط الدقة لمقاييس rouge أعلاه في جميع الموضوعات في مجموعة بيانات DUC 2002 هو (0.81 و 0.75 و 0.81) على التوالي عندما يكون ملخص الكلمة المرجعية 200 كلمة ؛ (0.76 و 0.69 و 0.76) على التوالي عندما يكون ملخص الكلمة المرجعية 400 كلمة ؛ و (0.78 و 0.72 و 0.78) على التوالي عندما يكون متوسط ملخص الكلمات المرجعية بين 200 و 400 كلمة. في حين أن النتائج F- لهذه المقاييس أعلاه في مجموعة البيانات الثانية لثلاثة مواضيع هي {(0.64،0.76 و 0.76) و (0.62،0.69 و 0.69) و (0.96،0.92،0.96)} على التوالي عند الكلمة المرجعية الملخص هو (390114 ، 518) كلمة على التوالي

جمهورية العراق

وزارة التعليم العالي والبحث العلمي

جامعة بابل - كلية تكنولوجيا المعلومات

قسم البرمجيات

تلخيص استخلاصي للموضوع بناءً على تحسين تجميع الجمل المتسلسلة المعجمية والشبكة العصبية

اطروحة مقدمة

الى مجلس كلية تكنولوجيا المعلومات - جامعة بابل وهي جزء من متطلبات نيل

درجة الدكتوراه فلسفة في تكنولوجيا المعلومات / برمجيات

من قبل

مروان بدران محمد رشيد

إشراف

أ.م. د. وفاء محمد سعيد حمزه

1442 هـ

2022 م