International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 15 (2023) pp. 212-226

Received: 19 Nov. 2022; Accepted:13 May, 2023; Published: 9 June, 2023

Intelligent Abstractive Text Summarization using Hybrid Word2Vec and Swin Transformer for Long Documents

Gitanjali Mishra [1], Nilamber Sethi [2], Agilandeeswari L[3], and Yu-Chen Hu

[1]Department of CSE, GIET University,
Gunupur, At – Gobriguda, Po- Kharling, Gunupur, Odisha 765022, India
gitamca3@gmail.com

[2] Department of CSE, GIET University,
Gunupur, At – Gobriguda, Po- Kharling, Gunupur, Odisha 765022, India
nilambar@giet.edu

[3]School of Information Technology and Engineering, Vellore Institute of Technology,
Tiruvalam Road, Vellore - 632014, Tamil Nadu, India.
agila.l@vit.ac.in

[4]Department of Computer Science and Information Management, Providence University,
Taichung 43301, Taiwan, Taiwan,
ychu@pu.edu.tw

***Abstract*: In this paper, a novel hybrid Swin-T transformer based automatic text summarization model is proposed for both short and long documents. It involves various phases namely data acquisition, preprocessing, Word to Vector conversion, semantic feature extraction using Swin-T transformer, clustering the similar sentences using K-medoid clustering and finally ranking the sentences and summary generation using Bi-directional Gated Recurrent Unit (Bi-GRU). This setup outperforms the existing state-of-the-art systems in terms of evaluation score named ROUGE score such as ROUGE 1, ROUGE 2 and ROUGE L for the short benchmark datasets as well as a long user created arXiv dataset.**

***Keywords*: Swin-T Transformer, K-Medoid, Bi-directional Gated Recurrent Unit, Semantic Feature Extraction, ROUGE, Word2Vector.**

I. Introduction

Nowadays, we adore rapid access to vast quantities of data from the online social media through tweets, news articles, blogs, reports, etc. However, the majority of the information accessed is unnecessary, unimportant, and might not accurately reflect the required information. This unnecessary junk make it difficult to filter and time consuming when a specific information is searched on an online social media. The automatic text summarization (ATS) capable of mining information that is useful but excludes material that is pointless and superfluous. Text summarization is a strategy for distilling lengthy texts into a clear, accurate summary without losing the general meaning or the information they contain. The goal of automatic text summarization is to reduce the length of documents' contents into shorter versions of them. Manual text summarization could be time-consuming and expensive. [1]. These hitches can be overwhelmed using an Automatic text summarization approach and facilitate to produce the key concepts in a portion of script contentedly [2]. The current expansion of non-structured textual data in the digital sphere necessitates the creation of automatic text summary technologies that make it simple for users to draw conclusions from them. Implementing summarization can make documents easier to read, save time spent looking up information again, and allow for the fitting of more information into a given space.

According to International Data Corporation (IDC), the amount of digital data that is transmitted globally each year will increase from 4.4 zettabytes in 2013 to 180 zettabytes in 2025. Because there is so much data floating around in the digital world, algorithms that can automatically condense longer texts and provide precise summaries that effectively convey the intended messages must be developed. Additionally, using text summarization shortens reading sessions, speeds up information research, and expands the amount of information that can fit in a given space.

Generally, text summarization techniques are classified into two types on the basis of output or summary generation namely, (i) extractive and (ii) abstractive [3]. As the name implies, an extractive summarization involves extracting only the keywords or key sentences from the text document without any modifications. The extractive summarization encompasses the following steps such as, text input and its intermediate representation, sentence score calculation and the extraction of sentences [4, 5]. On the other hand

abstractive summarization involves paraphrasing i.e., instead of extraction it dealt with word or sentence rephrasing and thereby generating the summary of new words [6, 7]. Most of the research has dealt with abstractive text summarizations as it seems to be the best way of generating a summary.

Based on the source documents size for summary generation, text summarization is classified as single-document, where a single document is given as input or multi-document, where a set of documents are given as input for producing the summary. This single document and multi-documents can be further defined with long documents [8] or short documents [9]. The extractive summarization techniques are performed using supervised [10], unsupervised [11] and hybrid [12] techniques. In supervised, the labelled data is used for training and thereby produces a summary, whereas, in an unsupervised systems, no labelled data is used for summary generation. Thus, from the study it is clear that extractive summarization is not only simple but also faster than the abstractive [13] with good accuracy [14] qualities of extractive summarization are measured using ROUGE Score, Precision, Recall and F1 Score.

Long document summarisation is an open problem in NLP. Here we review previous attempts on the problem and then we proceed to introduce our own modified Swin Transformer and Word2Vec for long documents summarization which attempt to use a Large Language Model based on the transformer architecture.

We are focusing on neural "abstractive" summarization of long documents. First we will go through some of the challenges in doing long-form document summarisation.

a) One of the biggest problems related to 'context'. A language model can only capture some amount of context. And the size of 'context' required for doing correct summarization is very large in case of long documents.

b) Another problem is the fundamental limit of popular language models, i.e. they only accept a maximum number of words as input. So long documents exceeding that maximum amount can't be directly fed into the model for inference.

c) Previous approaches that used RNN and LSTM based models suffered from the 'Long term dependence problem', so an 'abstractive' summary couldn't be generated for long documents.

Here, in this article, we are concentrating on the long document summarization.

The rest of the article is presented as follows: Section II delivers a detailed literature. The proposed methodology is demonstrated in section III. The Experimental Analysis section IV validates the proposed system through various short and long documents benchmark datasets. Finally, Conclusions were given in section V.

II. Literature Survey

This section provides an overview of the field of long text summarization, including the various techniques and models that have been developed to tackle the challenges of summarizing long texts. It covers both traditional methods, such as rule-based systems, and more recent deep learning techniques. It also discusses the current limitations of long text summarization models, such as information loss, inaccuracies, limited context, and difficulty in preserving coherence. Additionally, it highlights the evaluation metrics used in the field to assess the quality of generated summaries. The survey concludes with a discussion of future directions for research in long text summarization and the potential for further improvement of the current models. Overall, the survey provides a comprehensive overview of the current state of the art in long text summarization and its various applications. Sotudeh et al., proposed a Bi-directional LSTM for encoder, and LSTM for decoder. It uses two encoders for different parts of a document that easily captures different contexts. But, LSTMs have the long-term dependence problem. This system achieves the various Rouge scores as Rouge-1 = 52.47, Rouge-2 = 40.11, and Rouge-L = 51.39 [15] Su D et al., proposed a fine tuning BART model (Sequence to Sequence denoising autoencoder). Through the process of denoising, the corrupted text can be identified also the model learns which parts of the text are to be in context. As it is a denoising autoencoder, the right context that is learnt depends on the amount of corruption in the original text. The Rouge scores obtained by this model are Rouge-1 = 52.47, Rouge-2 = 40.11, and Rouge-L = 51.39 [16].

Ngamcharoen et al. suggested using overlapping words from the source text and the reference summary as the ground truth for a bi-directional long short-term memory model, with the first set representing words from the source text and the second set representing words from the reference summary. ROUGE-1 F1 is 0.0301, while ROUGE-2 F1 is 0.0140, according to the tests' findings on the ThaiSum dataset [17].

Cachola et al., uses a Transformer based denoising autoencoder BART pretrained on XSUM dataset helps in doing abstractive summarization. It has the same issue as discussed in the above case, about corruption of input corpus. The obtained Rouge scores are Rouge-1 = 43.8, Rouge-2 = 20.9, and Rouge-L = 35.5 [18]. S. Wang et al., introduced a two-phase approach towards long text summarization named EA- LTS. Extraction, involves the use of a hybrid sentence similarity measure, whereas the Abstraction, involves the construction of a recurrent neural network. The use of the hybrid sentence similarity measure, is said to result in improved accuracy and validity compared to existing state-of-the-art methods. The proposed recurrent neural network based encoder- decoder and attention mechanisms may not be suitable for all types of long texts and may need to be adjusted for different domains and genres. The obtained Rouge-1, and Rouge was 33.9 and 2.11 respectively [19].

Ahmed Magooda et al., proposed a Bidirectional Attention that can capture more semantics from the contexts. On the otherhand, there is a token limit to the model. This model attains a Rouge-1 = 39.36, Rouge-2 = 17.17, and Rouge-L = 26.78 [20].

Sarkhel et al., introduced a novel Multi level summarizer model. The biggest advantage is that the encoder layer kernels are handwritten according to the domain. The kernels are hardcoded so that means they can't be learnt. The Rouge scores attained are Rouge-1 = 45.99, Rouge-2 = 35.97, and Rouge-L = 40.89 [21].

Hernández-Castañeda, Ángel, et al proposed a sentence extraction-based query-oriented text summarization method is suggested. The text's most informative sentences are found and chosen to be included in the summary using the extractive

summarising approach. A collection of pertinent attributes is retrieved from the text in order to identify sentences that contain important information. The most informative sentences are more precisely recognised, the extracted sentence attributes are more relevant, and the quality of the resulting summary is improved [22].

In [23] authors used the BBC news dataset from Kaggle for their research project in an effort to as closely imitate the use of voice data as feasible. We attempt to address the summarization issue using the Bi-LSTM machine learning model. They also highlight the distinction between text summary at the sentence and paragraph levels. The findings of our model were then compared to those of both of these methods, and it was discovered that our model fared far better in the situation of sentence-level text summarization.

Yao, Kaichun, et al proposed a Abstractive text summarization works well with recurrent neural network-based sequence-to-sequence attentional models. This study models abstractive text summarization using a dual encoding paradigm. The suggested approach uses both primary and secondary encoders, unlike earlier works. The secondary encoder models word importance and creates finer encoding depending on input raw text and output text summary. The decoder uses the two-level encodings to provide a more diversified summary that reduces repetition for long sequence generation. Our dual encoding model outperforms previous methods on two demanding datasets [24].

In [25], Ma et al., introduced a Data mining in information retrieval and natural language processing has accelerated in the age of social networks, necessitating automatic text summarization. In social network summarization, pretrained word embedding and sequence to sequence models can extract important information with strong encoding. These models must now address long text dependencies and use latent topic mapping. T-BERTSum is a topic-aware extractive and abstractive summarization model based on Bidirectional Encoder Representations from Transformers. (BERTs). The suggested method may simultaneously infer topics and summarise social texts, improving on earlier methods. To generate the topic, the neural topic model (NTM) matches the embedded BERT representation with the encoded latent topic representation. Second, the transformer network learns long term dependencies for end-to-end topic inference and text summarization. Third, LSTM network layers are stacked on the extractive model to gather sequence timing information, and a gated network filters the effective information on the abstractive model. A two-stage extractive–abstractive model shares the information. TBERTSum employs pretrained external knowledge and topic mining to improve contextual representations compared to previous work. Our model generates consistent topics and delivers new state-of-the-art outcomes on CNN/Daily mail and XSum datasets.

Saraswathi, R. Vijaya, et al suggested a systems where the Common people find it harder to interpret consumer reviews on apps and websites as customers give varied product/service reviews. Despite the time involved, some people are too lazy to read reviews before making a decision. No one can read every review. Thus, a text summary model would streamline this process. A text summary removes irrelevant or unimportant information from a lengthy work. LSTM-based text summarizers automatically summarise reviews. Separate and vectorize input sentences. Material summaries reduce big texts while maintaining context. Readability is key. We want to build a model that summarises food reviews. This helps meal orderers learn about their dish [26].

[27] Internet text resources have increased the requirement for automatic document summarising technologies. However, summarising tools must be improved. This work presents Karcı summarizing, a unique method for extractive, generic text summarising. In a novel document summarising method, Karcı Entropy was applied. The suggested system requires no information source or training data. KUSH (called after its creators, Karcı, Uçkan, Seyyarer, and Hark) was added at the input text stage to maintain semantic consistency across phrases. The Karcı Entropy-based approach selects the most effective, generic, and informative sentences in a paragraph. Karcı Summarization was tested using open-access document text (DUC-2002, DUC-2004) datasets. RecallOriented Understudy for Gisting Evaluation measures measured Karcı Summarization's performance. (ROUGE). The suggested summarizer surpassed all state-of-the-art approaches for 200-word summaries in ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-W-1.2 measures. On the DUC-2002 dataset, the suggested summarizer outperformed the closest competitors by 6.4% for ROUGE-1 Recall. These results show that Karcı Summarization is promising and should interest scholars. We demonstrated excellent adoptability. KUSH text processing made the approach insensitive to disordered and missing texts.

[28]. Redundancy in extractive text summarization is its biggest drawback. Over the past two decades, many extractive text summarising approaches have been presented, but redundancy has been neglected. This work proposes a vector space model-based text summarising method that uses topic modelling and semantic measuring to determine the best summary of the text. Our main goal is to reduce summarization redundancy by using just phrases that indicate the most themes in the text source. By expressing phrases in a vector space
model and topic modeling, we obtain the document's subject vector. To improve efficiency, we use the semantic similarity measure to determine sentence relevance. We present two methods for creating topic vectors from a document: combined and individual. Evaluation results on two datasets reveal that both variants (Combined and Individual topic vector techniques) of the proposed method provide summaries closer to human-generated summaries than existing text summarising methods.

[29]. Text summary condenses papers while retaining their essential information. Although there are a few datasets created for summarization tasks, most of them lack human-generated goal summaries, which are crucial for summary generation and evaluation. The majority of text summarization research has focused on supervised learning solutions. Thus, our work produced an abstractive and extractive summarization dataset. This dataset includes academic papers, author abstracts, and two-size excerpts created by human readers in this research. To verify human extract manufacturing, the extracts were tested. On the suggested dataset, the extractive summarization problem was revisited. To provide summaries that were more informative, the feature vector was studied. To that purpose, a thorough syntactic feature space was created for the proposed dataset, and its impact on summary informativeness was examined. GloVe and word2vec embeddings also contain semantic

information. Finally, a long short-term memory-based neural network model used ensembled feature space to combine syntactic and semantic information. ROUGE metrics assessed model summaries and found that the suggested ensemble feature space greatly enhanced syntactic or semantic feature single-use. The suggested approach's summaries on ensembled features outperformed or matched state-of-the-art extractive summarising algorithms.

[30]. Semantic Role Labeling (SRL) and Explicit Semantic Analysis are combined in the research article "SRLESA-TextSum: A Text Summarization Method Based on SRL and Explicit Semantic Analysis" to create a novel method for text summarization (ESA). The suggested approach uses ESA to determine the semantic similarity between sentences and choose the most informative ones for the summary after first using SRL to determine the functions of words in the source text. The authors test their strategy using the DUC-2004 and DUC-2005 datasets, compare it to existing summarising strategies, and demonstrate that it performs better than the alternatives in terms of ROUGE scores and human evaluation. The authors draw the conclusion that their
approach points the way forward for fruitful future study in the area of automatic text summarization.

[31]. An innovative approach for extractive text summarising that combines fuzzy evolutionary algorithms and clustering techniques is put forward in the research article "An Strategy for Extractive Text Summarization utilising Fuzzy Evolutionary and Clustering Algorithms." The suggested method employs a clustering algorithm to organise these phrases into coherent and representative summaries after selecting the key sentences from the original material using a fuzzy genetic algorithm. The authors test their strategy using the DUC-2002 dataset and compare it to existing summarising methods, demonstrating that their method outperforms the others in terms of ROUGE scores and human judgement. The authors draw the conclusion that their approach offers a promising avenue for further study in the area of automatic text summarising and is applicable to real-world tasks like document and news summarization.

[32]. An overview of several text summarising strategies, including extractive, abstractive, and hybrid approaches, is given in the research article "A Review of Various Forms of Text Summarization with their Satellite Contents Based on Swarm Intelligence Optimization Algorithms." The usage of satellite content in the summarization process—such as keywords and named entities—is also covered by the authors. The research also investigates the application of swarm intelligence optimisation techniques, such as particle swarm
optimisation, bee colony optimisation, and ant colony optimisation, for enhancing text summarization efficiency. The authors offer a thorough analysis of the body of knowledge in the area of automatic text summarization and point out a number of unresolved issues and potential future paths. Overall, the article is a useful tool for academics and professionals who are interested in text synthesis.

[5]. SummCoder, an unsupervised framework for extracting text summarization based on deep auto-encoders, is proposed in the research article "SummCoder." A deep auto-encoder model initially represents source text sentences as dense vector embeddings. A big text corpus trains the model unsupervised. The authors then use a clustering method to group the embeddings and choose the most representative sentences from each cluster for the summary. Using the CNN/DailyMail and DUC-2002 datasets, the authors compare their summarization method to others and show that it achieves state-of-the-art ROUGE ratings and human evaluation. The authors conclude that their technique is promising for future study in automatic text summarising and beneficial for news and document summarization.

[33]. The research article "Fuzzy Evolutionary Cellular Learning Automata Model for Text Summarization" presents a novel approach for extracting text summarising using fuzzy logic, evolutionary algorithms, and CLAs. Cellular learning automata are used to determine the most significant nodes in the graph of source text phrases. A fuzzy evolutionary algorithm is used to choose the most informative phrases from the detected nodes and provide a summary. For the DUC-2002 dataset, the authors compare their summarization method to others and find that it beats others in ROUGE scores and human evaluation. The authors conclude that their technique
is promising for future study in automatic text summarising and beneficial for news and document summarization.

[34]. Mahak Gambhir and Vishal Gupta's "Latest automated text summarising techniques: a survey" reviews current methods. The authors begin by emphasising the relevance of automatic text summarization in current times and then briefly describing its progress. The study then discusses extraction-based, abstraction-based, and hybrid automated text summarising algorithms. The authors analyse the pros and cons of each strategy and give examples of approaches developed utilising each. The study explores computerised text summarizing assessment measures and problems. Lastly, the authors address potential approaches in automated text summarising.

[35]. In the research article "Hybrid Method for Text Summarization Based on Statistical and Semantic Treatment," a hybrid method for text summarization is suggested that combines statistical and semantic methods. The method identifies and groups phrases with similar semantic content using latent semantic analysis (LSA), and then utilises statistical techniques to rate the sentences and choose the most useful ones for the summary. The suggested method was tested against a variety of datasets and contrasted with previous summary methods, demonstrating its ability to generate excellent summaries while retaining key details from the original text. The authors draw the conclusion that their hybrid technique points the way towards fruitful future study in the area of automatic text summarization.

[36]. The Bidirectional and Auto-Regressive Transformers (BART) model is suggested as a unique approach for abstractive text summarising in the research article "Abstractive Text Summarization Using BART." The suggested method makes use of the sequence-to-sequence architecture of BART and pre-trained weights to produce summaries that are similar to those written by humans and that effectively convey the meaning of the input text. The authors demonstrate the superiority of their methodology in terms of ROUGE scores and human evaluation by evaluating it on the CNN/DailyMail and XSum datasets and comparing it to other cutting-edge summarising algorithms. The authors come to the conclusion that their approach is a significant development in
the area of abstractive text summarization and has the potential to be used in a variety of contexts. Tay et al.,

presented a detailed study on efficient transformers and its various applications namely text summarization, generation, question and answering etc., [37].
Recently, Joshi et al., proposed a new DeepSumm model that delas with extractive summarization using a sequence to sequence based networks [38], whereas, Bani-Almarjeh, M., & Kurdy, M. B. proposed an abstractive text summarization using RNN-based and transformer-based architectures for Arabic texts [39].

A. *Motivations and Contributions*

1) *Motivations:*

In conclusion, the field of long text summarization has seen significant advances in recent years, with the development of various techniques and models. However, the task of summarizing long texts remains challenging, and current models still face a number of limitations, such as information loss, inaccuracies, limited context, and difficulty in preserving coherence. Despite these limitations, the use of deep learning techniques has shown promise in overcoming some of these challenges and producing more accurate and effective summaries. Further research is needed to improve the quality of long text summarization models, especially in preserving the context, tone, and human interpretation present in the original text. Nevertheless, the ongoing progress in the field shows great potential for the future of long text summarization and its various applications in various domains.

2) *Contributions:*

The major contributions of our proposed model is summarized as:

1. A novel hybrid Swin-T Transformer and BiGRU have been introduced.
2. The semantic similarity and extraction of sentences is achieved using Swin – T Transformer. We tried this and succeeded in the initial attempt.
3. The word2Vec based word embedding are used for converting the given words to vector to make it suitable for Swin –T Transformer.
4. The K-Medoid clustering are used to cluster the sentences.
5. Finally, Bi-GRU are used to rank the sentences by assigning the score for the clustered sentences and produce summary from the top scored clustered sentences.

III. Proposed Methodology

The proposed methodology involves the following phases, Data Acquisition, Preprocessing, Semantic feature extraction, K-Medoid Clustering, Ranking and Summary Generation.

A. *Step 1: Data Acquisition*

1) *CNN / Daily Mail*

Just over 300,000 unique news stories written by journalists for CNN and the Daily Mail make up the English-language dataset, especially in its combined CNN and Daily Mail form. The CNN and Daily Mail datasets are combined to provide additional data examples for better training. The current version supports both extractive and abstractive summarization, despite the fact that the initial version was created for automated reading, comprehension, and abstractive question answering. As per https://huggingface.co/datasets/viewer/?dataset=cnn_dailymail&config=3.0.0, the proposed models also use the average count of the token for the articles and the highlights are 781 and 56, respectively. The 3 major splits of CNN/Daily Mail dataset are in terms of training, testing, and validation, and the same are projected for Version 3.0.0 in Table 1.

Table 1. Splitting of CNN/DailyMail Dataset (Version 3).

Dataset Splitting	**Each Split Instances**
Training	287113
Testing	11490
Validation	13368

2) *BBC News*

This dataset was made using a dataset for data categorization from the 2004–2005 work by D. Greene and P. Cunningham [40], which consist of 2225 documents from the BBC news website relating to the news of five topical divisions (http://mlg.ucd.ie/datasets/bbc.html).

3) *DUC*

Document Understanding Conference Datasets are generated by the National Institute of Standards and Technology (NIST). The DUC corpus from 2002, 2003, 2004, 2006, and 2007 are used for evaluation and its URL is given as http://www-nlpir.nist.gov/projects/duc/data.

4) *X-Sum Dataset*

Created by Narayan et al. at 2018, the X-Sum. The 226,711 Wayback archived BBC articles from 2010 to 2017 that make up the XSum dataset are all in the English language and span a wide range of topics, including news, politics, sports, weather, business, technology, science, health, families, entertainment, and the arts. Containing 226,711 in JSON file format.

5) *arXiv*

There are more than 2 million academic publications in eight subject areas in the arXiv preprint database's complete corpus as of the arXiv annual report 2021. Along with the LaTeX source files, we gather the articles' published PDF versions. For better analysis, the abstract of each paper is collected from the relevant arXiv abstract page.

Thus, these datasets can be given as an input of the form short or long documents namely D_{short} *or* D_{long}

B. *Step 2: Pre-processing*

The above selected D_{short} *or* D_{long} documents will undergo various pre-processing stages for better text summarization. It includes:

1. Segmentation is used to organize the sentences after it is extracted from the documents.
2. Tokenization is used to extract the words from each sentence, mainly to identify the structure of the character, such as date and time, number, punctuation, etc.
3. Contraction mapping: Contractions are the most common in any online documents which deals with contracting a word or a groups of words by dipping the letters and supplanting them with an apostrophe. The text summarization gets affected by these contractions because,
   (i) it won't be easily understood in their context by conventional systems as it subjective in use,
   (ii) It is computationally expensive as it drastically increases the dimensionality of the vectorized text. To resolve this, contraction mapping are applied to map each sentence to its expanded form [41].
4. Case conversion involves converting the entire text in the input D_{short} *or* D_{long} document are converted into a lower case to maintain uniformity throughout the text.
5. Stop Word Removal concentrates on removing the common words such as 'a', 'an', 'the' which won't add any information to the input text.
6. Removal of Links/Emails deals with removing the undesirable text namely hyperlinks, URLs, and email addresses that were commonly occurs in the news articles and blogs using regex expressions.
7. Lemmatization is a process which converts the given words to its root word using a spaCy's POS lemmatizer, which includes both the part-of-speech (POS) and lemma of a word for conversion [42] For example, the words "eat", "eating", "eaten", "ate" are converted to its root form namely "eat".

C. *Step 3: Word Embedding*

This stage is used to convert the given words to a vector form that are suitable for the Swin-T Transformer. It mainly focusses on extracting the semantics of the text. A list of numbers is formed using word embedding, which replaces each word in a word dictionary with that word's syntax and meaning. The BOW (Bag of Words) approach is the standard model for word embedding. This approach requires sparse and high-dimensional input data, which makes it challenging to capture semantic links between textual units. The Word2Vec algorithm transforms each word into a vector of float numbered list based on the associations between words in an embedded low-dimensional space, overcoming the limitations of BOW [43]. As opposed to using the explicit semantic relations offered by external lexical relations like WordNet, Word2Vec expresses the implicit semantic links. Word2Vec is a two-layer neural network that processes the text. The vectors produced by Word2Vec are feature vectors of words. It has two architectures: skip-gram and Continuous Bag of Words (CBOW). Given the context, CBOW predicts the current word, whereas, Skip-gram predicts the context for the given word.

D. *Step 4: Sentence Extraction*

The vector form of the words are then passed into a Swin-T Transformer to extract the feature vectors of the input sentences and calculate the scores for each sentences with the help of Attentions. There are four attentions used in Swin-T Transformer and it helps to calculate the scores easily. [44] introduces the Swin Transformers, which come in four different sizes: Tiny (Swin-T), Small (Swin-S), Big (Swin-B), and Large (Swin-L). Swin -T and Swin -S have the same C = 96 and layer numbers of 2, 2, 6, 2 and 2, 2, 18, 2 respectively as their hyperparameters. Layer numbers for Swin-B with C = 128 are 2, 2, 18, 2, and for Swin-L with C = 192 are 2, 2, 18, 2. C is the channel number of the hidden layers in stage 1. For easier comprehension, the two successive Swin Transformer blocks' miniature version (Swin-T) from [44] is depicted in Fig. 1.

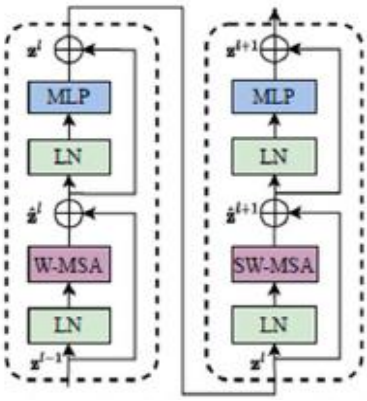


Figure 1. Swin Transformer Block [44]

(1) The Attention

As the model generates words, it has the ability to reference or 'attend' to words that are relevant to the generated word. And the model learns which words to attend to while training with backpropagation. This window of words to reference is what we call 'context'.

(2) Self-Attention:

Self-attention enables the model to associate each individual word in the input to other words in the input. And which are the important words to attend to, or what is the important 'context' for some words is learnt through backpropagation. Self-attention technique is applied in the "Encoder". For example in the sentence "The agreement on the European economic area was signed in 1942", When looking at the word "signed" which words should the model use as context : 'European', 'economic' , 'area' or "agreement"

(3) Encoder

The encoder maps an input sequence into a continuous vector representation that holds all learnt information of that input. The input sequence is a vector representation of each word. We get this vector by using word embedding's. The main part of the encoder is multiple self-attention units, they capture the attention weights (i.e. how much attention should each words attend to or refer other words). Usually there are many encoders where each one has the opportunity to learn different attention representations from different contexts.

(4) Decoder

The decoder then uses that encoded representation and feeds it the previous output as it gradually produces a single word. The decoder will be able to focus on the appropriate words thanks to the attention weights learned during the encoding stage (or the right context) when it is outputting new words.

E. K-Medoid Clustering

The extracted sentences from Swin-T Transformer is given as input to the K-Medoid clustering in order to group based on the similarity scores. All sentences are grouped together around a centroid and ranked according to Euclidean distance by the K-Medoid algorithm [45]. It is less susceptible to outliers and noise and has a faster rate of convergence than TextRank [46] and K-Means [47] clustering within discrete steps.

F. Ranking using Bi-directional Gated Recurrent Unit (Bi-GRU)

The Bi-GRU based encoder – decoder model is followed by the K-Medoid clustering to generate the summary based on the top rank of each clusters. Once the clusters are formed, it is then sent to the Bi-directional encoder of GRU followed by the decoder to rank the sentences in each clusters after removing the similarities. These top ranked sentences will be used in the summary generation process and thus the summary is generated.

G. Summary Generation

Finally, the output of Bi-GRU is then backpropagates to the Swin-T transformer to extract the perfect semantics and repeats the k-medoid and Bi-GRU stages until the meaningful summary generation happens with a maximum ROUGE score.

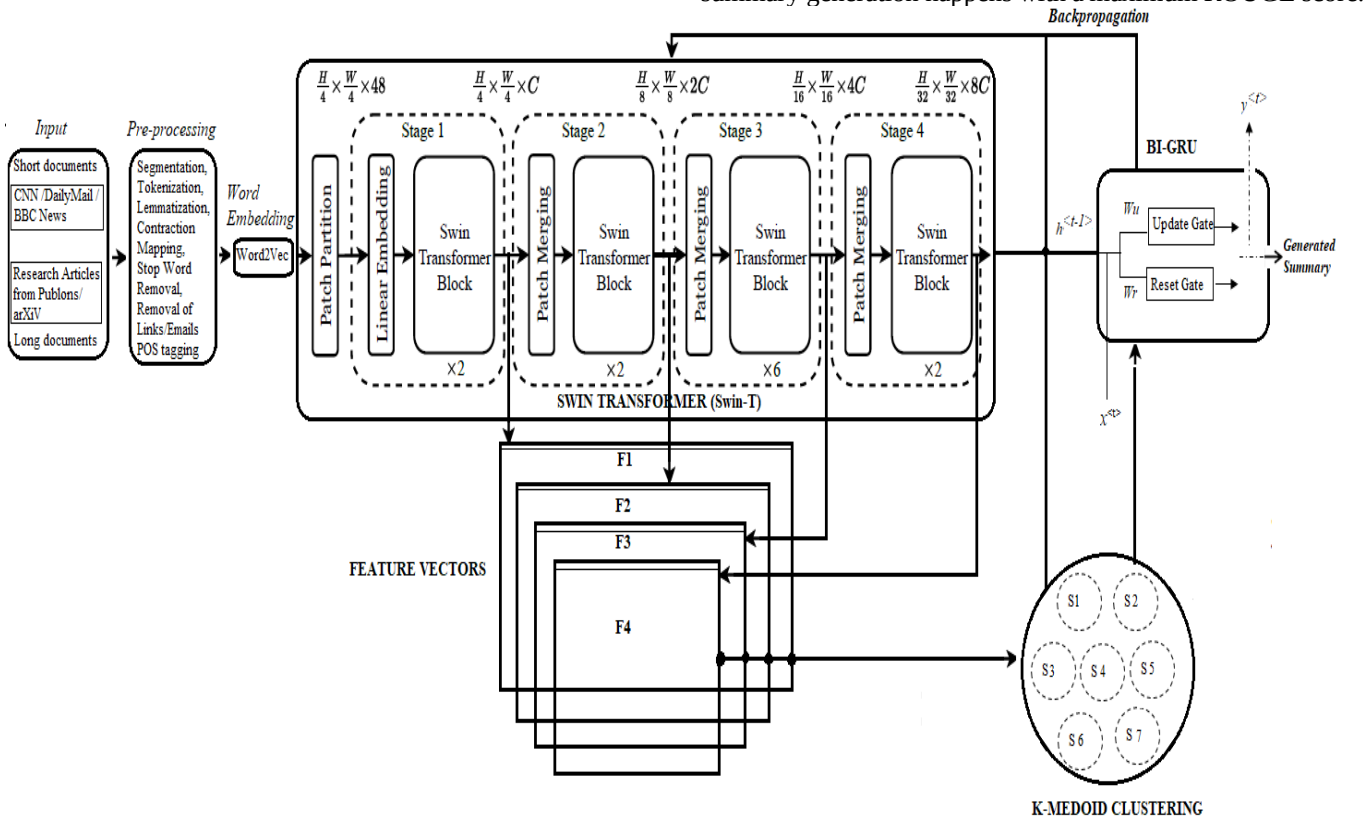


Figure 2. Proposed Text Summarization Model

IV. Experimental Analysis

On the datasets supplied in Section III A, the effectiveness of the proposed framework was evaluated. The Swin-T transformer, and K-Medoid clustering were implemented in Python using Scikit-learn. As previously mentioned, all data points were used at once without any training, testing, or validation divides because the suggested approach is unsupervised. The Google Collaboratory (CPU) [48] was used for all testing, and the average processing time from preprocessing to summary production was 110,000 seconds.

A. Qualitative Analysis

Comparing the automatically generated summaries to the summaries created by humans, or "human-generated summaries," allows for a qualitative study of the suggested summarization model. The sample text from the BBC News dataset is displayed in Fig. 2, and Fig. 3 shows both the generated summary and the ground truth summary. The Fig.3 proves that proposed summarization model works well for both short and long documents, respectively. The following sample texts from BBC News are considered for validating the proposed system.

Ink helps drive democracy in Asia

The Kyrgyz Republic, a small, mountainous state of the former Soviet republic, is using invisible ink and ultraviolet readers in the country's elections as part of a drive to prevent multiple voting. This new technology is causing both worries and guarded optimism among different sectors of the population. In an effort to live up to its reputation in the 1990s as "an island of democracy", the Kyrgyz President, Askar Akaev, pushed through the law requiring the use of ink during the upcoming Parliamentary and Presidential elections. The US government agreed to fund all expenses associated with this decision. The Kyrgyz Republic is seen by many experts as backsliding from the high point it reached in the mid-1990s with a hastily pushed through referendum in 2003, reducing the legislative branch to one chamber with 75 deputies. The use of ink is only one part of a general effort to show commitment towards more open elections - the German Embassy, the Soros Foundation and the Kyrgyz government have all contributed to purchase transparent ballot boxes. The actual technology behind the ink is not that complicated. The ink is sprayed on a person's left thumb. It dries and is not visible under normal light. However, the presence of ultraviolet light (of the kind used to verify money) causes the ink to glow with a neon yellow light. At the entrance to each polling station, one election official will scan voter's fingers with UV lamp before allowing them to enter, and every voter will have his/her left thumb sprayed with ink before receiving the ballot. If the ink shows under the UV light the voter will not be allowed to enter the polling station. Likewise, any voter who refuses to be inked will not receive the ballot. These elections are assuming even greater significance because of two large factors - the upcoming parliamentary elections are a prelude to a potentially regime changing presidential election in the Autumn as well as the echo of recent elections in other former Soviet Republics, notably Ukraine and Georgia. The use of ink has been controversial - especially among groups perceived to be pro-government. Widely circulated articles compared the use of ink to the rural practice of marking sheep - a still common metaphor in this primarily agricultural society.

The author of one such article began a petition drive against the use of the ink. The greatest part of the opposition to ink has often been sheer ignorance. Local newspapers have carried stories that the ink is harmful, radioactive or even that the ultraviolet readers may cause health problems. Others, such as the aggressively middle of the road, Coalition of Non-governmental Organizations, have lauded the move as an important step forward. This type of ink has been used in many elections in the world, in countries as varied as Serbia, South Africa, Indonesia and Turkey. The other common type of ink in elections is indelible visible ink - but as the elections in Afghanistan showed, improper use of this type of ink can cause additional problems. The use of "invisible" ink is not without its own problems. In most elections, numerous rumors have spread about it. In Serbia, for example, both Christian and Islamic leaders assured their populations that its use was not contrary to religion. Other rumours are associated with how to remove the ink - various soft drinks, solvents and cleaning products are put forward. However, in reality, the ink is very effective at getting under the cuticle of the thumb and difficult to wash off. The ink stays on the finger for at least 72 hours and for up to a week. The use of ink and readers by itself is not a panacea for election ills. The passage of the inking law is, nevertheless, a clear step forward towards free and fair elections." The country's widely watched parliamentary elections are scheduled for 27 February. David Mikosz works for the IFES, an international, non-profit organisation that supports the building of democratic societies.

(a)

Ad sales boost Time Warner profit

Quarterly profits at US media giant TimeWarner jumped 76% to $1.13bn (£600m) for the three months to December, from $639m year-earlier. The firm, which is now one of the biggest investors in Google, benefited from sales of high-speed internet connections and higher advert sales. TimeWarner said fourth quarter sales rose 2% to $11.1bn from $10.9bn. Its profits were buoyed by one-off gains which offset a profit dip at Warner Bros, and less users for AOL. Time Warner said on Friday that it now owns 8% of search-engine Google. But its own internet business, AOL, had has mixed fortunes. It lost 464,000 subscribers in the fourth quarter profits were lower than in the preceding three quarters. However, the company said AOL's underlying profit before exceptional items rose 8% on the back of stronger internet advertising revenues. It hopes to increase subscribers by offering the online service free to TimeWarner internet customers and will try to sign up AOL's existing customers for high-speed broadband. TimeWarner also has to restate 2000 and 2003 results following a probe by the US Securities Exchange Commission (SEC), which is close to concluding. Time Warner's fourth quarter profits were slightly better than analysts' expectations. But its film division saw profits slump 27% to $284m, helped by box-office flops Alexander and Catwoman, a sharp contrast to year-earlier, when the third and final film in the Lord of the Rings trilogy boosted results. For the full-year, TimeWarner posted a profit of $3.36bn, up 27% from its 2003 performance, while revenues grew 6.4% to $42.09bn. "Our financial performance was strong, meeting or exceeding all of our full-year objectives and greatly enhancing our flexibility," chairman and chief executive Richard Parsons said. For 2005, TimeWarner is projecting operating earnings growth of around 5%, and also expects higher revenue and wider profit margins. TimeWarner is to restate its accounts as part of efforts to resolve an inquiry into AOL by US market regulators. It has already offered to pay $300m to settle charges, in a deal that is under review by the SEC. The company said it was unable to estimate the amount it needed to set aside for legal reserves, which it previously set at $500m. It intends to adjust the way it accounts for a deal with German music publisher Bertelsmann's purchase of a stake in AOL Europe, which it had reported as advertising revenue. It will now book the sale of its stake in AOL Europe as a loss on the value of that stake.

(b)

Gallery unveils interactive tree

A Christmas tree that can receive text messages has been unveiled at London's Tate Britain art gallery. The spruce has an antenna which can receive Bluetooth texts sent by visitors to the Tate. The messages will be "unwrapped" by sculptor Richard Wentworth, who is responsible for decorating the tree with broken plates and light bulbs. It is the 17th year that the gallery has invited an artist to dress their Christmas tree. Artists who have decorated the Tate tree in previous years include Tracey Emin in 2002. The plain green Norway spruce is displayed in the gallery's foyer. Its light bulb adornments are dimmed, ordinary domestic ones joined together with string. The plates decorating the branches will be auctioned off for the children's charity ArtWorks. Wentworth worked as an assistant to sculptor Henry Moore in the late 1960s. His reputation as a sculptor grew in the 1980s, while he has been one of the most influential teachers during the last two decades. Wentworth is also known for his photography of mundane, everyday subjects such as a cigarette packet jammed under the wonky leg of a table.

(c)

Labour plans maternity pay rise

Maternity pay for new mothers is to rise by £1,400 as part of new proposals announced by the Trade and Industry Secretary Patricia Hewitt. It would mean paid leave would be increased to nine months by 2007, Ms Hewitt told GMTV's Sunday programme. Other plans include letting maternity pay be given to fathers and extending rights to parents of older children. The Tories dismissed the maternity pay plan as "desperate", while the Liberal Democrats said it was misdirected. Ms Hewitt said: "We have already doubled the length of maternity pay, it was 13 weeks when we were elected, we have already taken it up to 26 weeks. "We are going to extend the pay to nine months by 2007 and the aim is to get it right up to the full 12 months by the end of the next Parliament." She said new mothers were already entitled to 12 months leave, but that many women could not take it as only six of those months were paid. "We have made a firm commitment. We will definitely extend the maternity pay, from the six months where it now is to nine months, that's the extra £1,400." She said ministers would consult on other proposals that could see fathers being allowed to take some of their partner's maternity pay or leave period, or extending the rights of flexible working to carers or parents of older children. The Shadow Secretary of State for the Family, Theresa May, said: "These plans were announced by Gordon Brown in his pre-budget review in December and Tony Blair is now recycling it in his desperate bid to win back women voters." She said the Conservatives would announce their proposals closer to the General Election. Liberal Democrat spokeswoman for women Sandra Gidley said: "While mothers would welcome any extra maternity pay the Liberal Democrats feel this money is being misdirected." She said her party would boost maternity pay in the first six months to allow more women to stay at home in that time. Ms Hewitt also stressed the plans would be paid for by taxpayers, not employers. But David Frost, director general of the British Chambers of Commerce, warned that many small firms could be "crippled" by the move. "While the majority of any salary costs may be covered by the government's statutory pay, recruitment costs, advertising costs, retraining costs and the strain on the company will not be," he said. Further details of the government's plans will be outlined on Monday. New mothers are currently entitled to 90% of average earnings for the first six weeks after giving birth, followed by £102.80 a week until the baby is six months old.

(d)

Claxton hunting first major medal

British hurdler Sarah Claxton is confident she can win her first major medal at next month's European Indoor Championships in Madrid. The 25-year-old has already smashed the British record over 60m hurdles twice this season, setting a new mark of 7.96 seconds to win the AAAs title. "I am quite confident," said Claxton. "But I take each race as it comes. "As long as I keep up my training but not do too much I think there is a chance of a medal." Claxton has won the national 60m hurdles title for the past three years but has struggled to translate her domestic success to the international stage. Now, the Scotland-born athlete owns the equal fifth-fastest time in the world this year. And at last week's Birmingham Grand Prix, Claxton left European medal favourite Russian Irina Shevchenko trailing in sixth spot. For the first time, Claxton has only been preparing for a campaign over the hurdles - which could explain her leap in form. In previous seasons, the 25-year-old also contested the long jump but since moving from Colchester to London she has re-focused her attentions. Claxton will see if her new training regime pays dividends at the European Indoors which take place on 5-6 March.

(e)

Figure 2. BBC News Dataset(a) Source Document – Tech, (b) Source Document – Business, (c) Source Document – Entertainment, (d) Source Document – Politics, and (e) Source Document - Sport

Gold Summary	Generated Summary
Source Document 1 – Tech	
The other common type of ink in elections is indelible visible ink - but as the elections in Afghanistan showed, improper use of this type of ink can cause additional problems. The use of ink and readers by itself is not a panacea for election ills. The use of "invisible" ink is not without its own problems. The use of ink is only one part of a general effort to show commitment towards more open elections - the German Embassy, the Soros Foundation and the Kyrgyz government have all contributed to purchase transparent ballot boxes. The author of one such article began a petition drive against the use of the ink. The use of ink has been controversial - especially among groups perceived to be pro-government. In an effort to live up to its reputation in the 1990s as "an island of democracy", the Kyrgyz President, Askar Akaev, pushed through the law requiring the use of ink during the upcoming Parliamentary and Presidential elections. At the entrance to each polling station, one election official will scan voter's fingers with UV lamp before allowing them to enter, and every voter will have his/her left thumb sprayed with ink before receiving the ballot. The ink is sprayed on a person's left thumb. If the ink shows under the UV light the voter will not be allowed to enter the polling station. The actual technology behind the ink is not that complicated. The Kyrgyz Republic, a small, mountainous state of the former Soviet republic, is using invisible ink and ultraviolet readers in the country's elections as part of a drive to prevent multiple voting. This type of ink has been used in many elections in the world, in countries as varied as Serbia, South Africa, Indonesia and Turkey.	The Kyrgyz Republic is using invisible ink and ultraviolet readers in the country's elections as part of a drive to prevent multiple voting. This new technology is causing both worries and guarded optimism among different sectors of the population. In an effort to live up to its reputation in the 1990s as "an island of democracy", the Kyrgyz President, Askar Akaev, pushed through the law requiring the use of ink during the upcoming Parliamentary and Presidential elections. The US government agreed to fund all expenses associated with this decision. The Kyrgyz Republic is seen by many experts as backsliding from the high point it reached in the mid-1990s with a hastily pushed through referendum in 2003, reducing the legislative branch to one chamber with 75 deputies. The use of ink is only one part of a general effort to show commitment towards more open elections - the German Embassy, the Soros Foundation and the Kyrgyz government have all contributed to purchase transparent ballot boxes. The actual technology behind the ink is not that complicated. The ink is sprayed on a person's left thumb. It dries and is not visible under normal light. However, the presence of ultraviolet light (of the kind used to verify money) causes the ink to glow with a neon yellow light. At the entrance to each polling station, one election official will scan voter's fingers with UV lamp before allowing them to enter, and every voter will have his/her left thumb sprayed with ink before receiving the ballot.
Gold Summary	**Generated Summary**
Source Document 2 – Business	
TimeWarner said fourth quarter sales rose 2% to $11.1bn from $10.9bn.For the full-year, TimeWarner posted a profit of $3.36bn, up 27% from its 2003 performance, while revenues grew 6.4% to $42.09bn.Quarterly profits at US media giant TimeWarner jumped 76% to $1.13bn (£600m) for the three months to December, from $639m year-earlier. However, the company said AOL's underlying profit before exceptional items rose 8% on the back of stronger internet advertising revenues. Its profits were buoyed by one-off gains which offset a profit dip at Warner Bros, and less users for AOL. For 2005, TimeWarner is projecting operating earnings growth of around 5%, and also expects higher revenue and wider profit margins. It lost 464,000 subscribers in the fourth quarter profits were lower than in the preceding three quarters. Time Warner's fourth quarter profits were slightly better than analysts' expectations.	Time Warner's quarterly profits jumped 76% to $1.13bn for the three months to December, from $639m year-earlier, buoyed by sales of high-speed internet connections and higher advert sales. One-off gains offset a profit dip at Warner Bros, and less users for AOL. Time Warner said on Friday that it now owns 8% of search-engine Google. But its own internet business, AOL, had has mixed fortunes. It lost 464,000 subscribers in the fourth quarter profits were lower than in the preceding three quarters. However, the company said AOL's underlying profit before exceptional items rose 8% on the back of stronger internet advertising revenues. For the full-year, TimeWarner posted a profit of $3.36bn, up 27% from its 2003 performance, while revenues grew 6.4% to $42.09bn. TimeWarner is to restate its accounts as part of efforts to resolve an inquiry into AOL by US market regulators. It has already offered to pay $300m to settle charges, in a deal that is under review by the SEC. The company said it was unable to estimate the amount it needed to set aside for legal reserves, which it previously set at $500m.

Gold Summary	Generated Summary
Source Document 3 – Entertainment	
The messages will be "unwrapped" by sculptor Richard Wentworth, who is responsible for decorating the tree with broken plates and light bulbs. A Christmas tree that can receive text messages has been unveiled at London's Tate Britain art gallery. It is the 17th year that the gallery has invited an artist to dress their Christmas tree. The spruce has an antenna which can receive Bluetooth texts sent by visitors to the Tate. His reputation as a sculptor grew in the 1980s, while he has been one of the most influential teachers during the last two decades.	The text describes the unveiling of an interactive Christmas tree at the Tate Britain art gallery in London. The tree, which is decorated with broken plates and light bulbs, can receive text messages from visitors via Bluetooth. The artist responsible for the tree, Richard Wentworth, will "unwrap" the messages and auction off the plates for a children's charity. Wentworth is a well-known sculptor and teacher who is known for his photography of mundane, everyday subjects.
Gold Summary	**Generated Summary**
Source Document 4 – Politics	
She said her party would boost maternity pay in the first six months to allow more women to stay at home in that time. .She said new mothers were already entitled to 12 months leave, but that many women could not take it as only six of those months were paid. The Tories dismissed the maternity pay plan as "desperate", while the Liberal Democrats said it was misdirected. She said ministers would consult on other proposals that could see fathers being allowed to take some of their partner's maternity pay or leave period, or extending the rights of flexible working to carers or parents of older children. Liberal Democrat spokeswoman for women Sandra Gidley said: "While mothers would welcome any extra maternity pay the Liberal Democrats feel this money is being misdirected". We will definitely extend the maternity pay, from the six months where it now is to nine months, that's the extra £1,400."Ms Hewitt said: "We have already doubled the length of maternity pay, it was 13 weeks when we were elected, we have already taken it up to 26 weeks". Other plans include letting maternity pay be given to fathers and extending rights to parents of older children.	The UK Trade and Industry Secretary Patricia Hewitt has announced plans to raise maternity pay by £1,400, which would mean paid leave would be increased to nine months by 2007. The Tories have dismissed the maternity pay plan as "desperate", while the Liberal Democrats said it was misdirected. Ms Hewitt said that the government would consult on other proposals that could see fathers being allowed to take some of their partner's maternity pay or leave period, or extending the rights of flexible working to carers or parents of older children. The Shadow Secretary of State for the Family, Theresa May, said that the plans were announced by Gordon Brown in his pre-budget review in December and that Tony Blair is now recycling it in his desperate bid to win back women voters. Liberal Democrat spokeswoman for women Sandra Gidley said that while mothers would welcome any extra maternity pay, the Liberal Democrats feel this money is being misdirected. She said her party would boost maternity pay in the first six months to allow more women to stay at home in that time. Ms Hewitt also stressed the plans would be paid for by taxpayers, not employers. But David Frost, director general of the British Chambers of Commerce, warned that many small firms could be "crippled" by the move.
Gold Summary	**Generated Summary**
Source Document 5 – Sport	
For the first time, Claxton has only been preparing for a campaign over the hurdles - which could explain her leap in form. Claxton has won the national 60m hurdles title for the past three years but has struggled to translate her domestic success to the international stage. British hurdler Sarah Claxton is confident she can win her first major medal at next month's European Indoor Championships in Madrid Claxton will see if her new training regime pays dividends at the European Indoors which take place on 5-6 March."I am quite confident," said Claxton.	Sarah Claxton is a British hurdler who is confident she can win her first major medal at next month's European Indoor Championships in Madrid. Claxton has already smashed the British record over 60m hurdles twice this season, setting a new mark of 7.96 seconds to win the AAAs title. Claxton has won the national 60m hurdles title for the past three years but has struggled to translate her domestic success to the international stage. Now, the Scotland-born athlete owns the equal fifth-fastest time in the world this year. And at last week's Birmingham Grand Prix, Claxton left European medal favourite Russian Irina Shevchenko trailing in sixth spot. For the first time, Claxton has only been preparing for a campaign over the hurdles - which could explain her leap in form. In previous seasons, the 25-year-old also contested the long jump but since moving from Colchester to London she has re-focused her attentions. Claxton will see if her new training regime pays dividends at the European Indoors which take place on 5-6 March.

Figure 3. Summary for BBC News Dataset(a) Source Document – Tech, (b) Source Document – Business, (c) Source Document – Entertainment, (d) Source Document – Politics, and (e) Source Document - Sport

Introduction – Abstract Pair	One of the leading causes of cancer death for both men and women is the lung cancer. The best way to improve the patient's chances for survival is the early detection of potentially cancerous cells. But, the conventional systems fails to segment the cancerous cells of various types namely, well-circumscribed, juxta-pleural, juxta-vascular and pleural-tail at its early stage (i.e., less than 3 mm) that leads to less classification accuracy. It is also noted that none of the existing systems achieved accuracy more than 98%. In this paper, we propose an optimal diagnosis system not only for early detection of lung cancer nodules and also to improve the accuracy in Fog computing environment. The Fog environment is used for storage of the high volume CT scanned images to achieve high privacy, low latency and mobility support. In our approach, for the accurate segmentation of Region of Interest (ROI), the hybrid technique namely Fuzzy C-Means (FCM) and region growing segmentation algorithms are used. Then, the important features of the nodule of interest such as geometric, texture and statistical or intensity features are extracted. From the above extracted features, the optimal features used for the classification of lung cancer are identified using the Cuckoo search optimization algorithm. Finally, the SVM classifier is trained using these optimal features, which in turn helps us to classify the lung cancer as either of type benign or malignant. The accuracy of the proposed system is tested using Early Lung Cancer Action Program (ELCAP) public database CT lung images. The total sensitivity and specificity attained in our system for the above said database are 98.13 and 98.79% respectively. This results in a mean accuracy of 98.51% for training and testing in a sample of 103 nodules occurring in 50 exams. The rate of false positives per exam was 0.109. Also, a high receiver operating characteristic (ROC) of 0.9962 has been achieved Nowadays, the usage of Fog computing platforms over the cloud computing environment is much appreciated for its efficiency and security (Rodrigo Roman et al. 2016; Yi et al. 2015). Cancer, in medical terms is known as malignant neoplasm, and there are more than 200 different types of cancer which may affect humans. According to American Cancer Society, lung cancer is the top killer cancer in both men and women in the United States of America (USA). During 2014, an estimated 224,210 new cases and 159,260 deaths due to lung cancer were predicted in the USA (Cancer facts and Figures 2014). It causes more deaths per year than the next four leading causes of cancer namely, prostate, pancreas, colon and breast. Based on the cellular characteristics lung cancer can be divided into two groups: Non-small lung cancer and small cell lung cancer. In general there are four stages of lung cancer, from I through IV. These stages are based on tumor size and tumor lymph node location (Lemjabbar-Alaoui et al. 2015). Only about 10-16% of patients survive for five or more years after the lung cancer diagnosis (Jemal et al. 2005; Micheli et al. 2003). This is mainly due to the detection of lung cancer cells at its middle (III) or at the advanced stage (IV) and thus making the treatments ineffective (Motohiro et al. 2002). The major radiographic indicators for the lung cancer are the pulmonary nodules, focal densities with diameter from 3 mm to 3 cm (Girvin and Ko 2008; El-Baz 2013). The only way to improve the patient's survival rate is through early detection of cancerous pulmonary nodules (Henschke et al. 1999; Sluimer et al. 2006). The pulmonary nodules that are typical shadows of pathological changes of lung cancer can be detected perfectly using the CT even at early stages when compared to the chest X-ray images (Prokop and Galanshi 2003; Hoffman and McLennan 1997; Zhao et al. 1999; Aubry et al. 2007). On the contrary, when compared to the diagnosis using chest X-ray images, the CT (I-ELCAP 2006) may exhaust radiologists due to the examination of more images for a single patient (for example, at least over 30 images per patient). This exhaustion of radiologists and their physical tiredness may lead to wrong diagnosis. Thus, the accurate manual assessment is time-consuming and highly inaccurate (Pu et al. 2009). To avoid this wrong diagnosis and to make the radiologists convenient, computer-aided diagnosis (CAD) systems have been developed (Okumura et al. 1998). It is a new technology for the automated pulmonary nodules segmentation. The accurate segmentation is essential for the exact assessment of nodule growth and to classify the malignant pulmonary nodules (cancerous cell) from the benign ones (non-cancerous cell) (Da Silva et al. 2010). The pulmonary nodules are categorized into four classes based on their position in the lung and proximity to other structures namely: well-circumscribed, juxta-vascular, juxta-pleural and pleural-tail nodules. This work aims at detecting all these lung nodules accurately through CT images. Our approach begins with an image pre-processing module to reduce the noises present on the low dose CT images using median filtering technique followed by background removal and thorax extraction. Second, in order to extract the lung region from its surroundings, a hybrid segmentation method is used. The proposed hybrid segmentation technique uses FCM algorithm for clustering and region growing method for eliminating the unwanted clusters, thus separating the exact lung regions. Third, by repeating region growing on the segmented lung region one can determine the nodule of interest accurately which leads to better classification accuracy. Fourth, in the feature extraction stage, the set of geometrical, texture and intensity or statistical features are computed for the identified nodule. Finally, in order to increase the accuracy the optimal features are determined from the calculated features for the segmented lung nodules using CS optimization algorithm and the same is trained by SVM classifier to classify the nodules as benign or malignant. Throughout this paper, the term nodule and tumor are used interchangeably.

Ground Truth Summary	Lung cancer is the most important cause of cancer death for both men and women. The early detection of lung cancer is very important to increase a patient's survival time. This paper proposes an Optimized Diagnosis System for early detection of lung cancer nodules from the chest Computer Tomography (CT) images using Cuckoo search algorithm and Support Vector Machine (SVM) classifier. In general, the complete diagnosis process involves various stages as: (i) pre-processing of images, (ii) segmentation of lung region from its surroundings, (iii) extracting nodules of interest and then determining the associated features which may be vital and (iv) Finally, classify the nodule into either benign or malignant. In this approach, for the accurate segmentation of Region of Interest (ROI), the hybrid technique namely Fuzzy C-Means (FCM) and region growing segmentation algorithms are used. Then the important features of the nodule of interest such as shape, statistical and texture are extracted. From the above extracted features, the optimized features used for the classification of lung cancer are identified using the cuckoo search optimization algorithm. Finally, the Support Vector Machine (SVM) classifier is trained using these optimized features, which in turn helps us to classify the lung cancer of type benign or malignant. The accuracy of the proposed system is validated using ELCAP public database CT lung images. The total sensitivity and specificity attained in our system for the above said database are 98.99% and 99.09% respectively. The False Positive Rate (FPR) of our system is 0.910 and the False Negative Rate (FNR) is 0.10. The overall accuracy rate of our proposed system is about 99.03 %, which is high enough when compared to the existing systems.
Generated Summary	The proposed system emphasizes on the accurate Region of Interest (ROI) segmentation using Fuzzy C-Means (FCM) and region growing algorithms. Then, the features are extracted through geometric, texture and statistical or intensity features. Among these extracted features, the optimal features are determined using the Cuckoo search optimization algorithm. Lastly, these optimal features are used to train the SVM classifier to classify the lung cancer as benign or malignant.

Figure 3. Long Document Summary

B. Quantitative Analysis

The top *'n'* sentences from each cluster are simply concatenated to create the final summary. The Recall Oriented-Understudy for Gisting Evaluation (ROUGE) measure [49] is employed for evaluation purposes. These measures include ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S. For ease of comparison with recent comparable research, the F-1 scores of ROUGE-1, ROUGE-2, and ROUGEL metrics are employed in this work. The n-gram recall between a candidate summary and a collection of reference summaries is calculated using ROUGE-N. The recall and precision numbers are used to calculate the F-1 score for ROUGE-N, which may be mathematically represented in equation

$$ROUGE - N = \frac{\sum_{S \in \{Ground\ truth\ Summaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{Ground\ truth\ Summaries\}} \sum_{gram_n \in S} Count\ (gram_n)} \quad (1)$$

where n *represents the n-gram's length*
$gram_n$ is the maximum co-occurring n-grams in the generated summary.
$Count_{match}(gram_n)$ is the maximum number of co-occurring n-grams in a set of reference summaries, and
Count is the total n-grams in the reference summaries.

Table 3. ROUGE Score of the Proposed Summarization model on the benchmark dataset (Section III A).

Dataset	ROUGE 1	ROUGE 2	ROUGE L
CNN/ DailyMail	56.45	29.58	53.23
BBC News	54.21	28.14	50.55
DUC 2002	56.86	30.21	54.85
DUC 2003	55.94	29.19	53.23

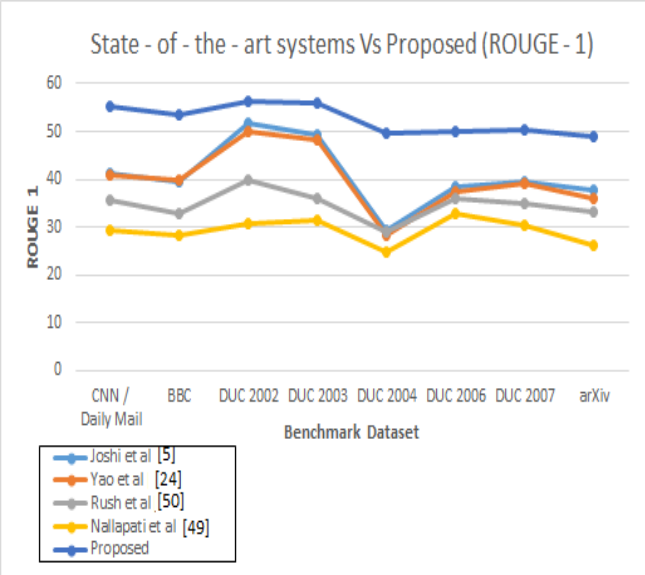
DUC 2004	50.23	24.54	48.21
DUC 2006	51.28	25.25	47.99
DUC 2007	50.56	26.45	49.56
X-Sum	51.47	29.32	51.27
arXiv	48.94	24.27	47.65

Table 4. ROUGE Score of the Proposed Summarization model on the BBC News Dataset as in Figure. 2.

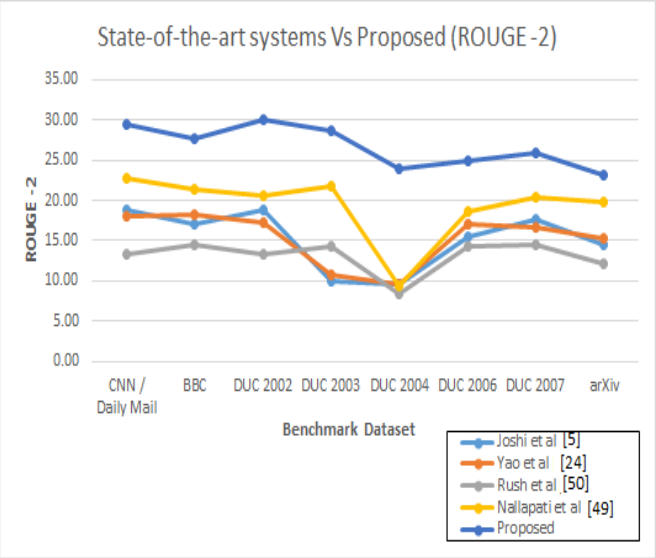
Dataset (BBC News)	**ROUGE 1**	**ROUGE 2**	**ROUGE L**
Source Doc.1	53.25	27.46	51.44
Source Doc.2	54.51	28.14	51.59
Source Doc.3	52.86	30.21	53.54
Source Doc.4	54.94	27.92	52.27
Source Doc.5	55.23	26.46	48.15

C. Comparative Analysis

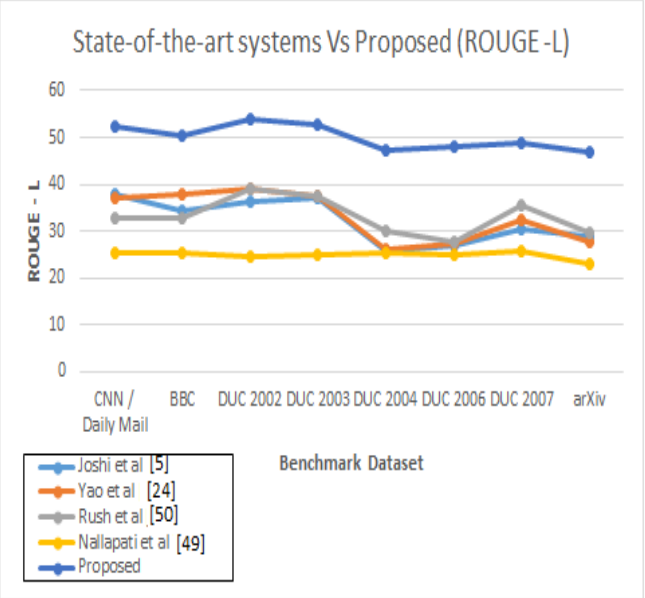
For a fair comparison, we have tested our model with the state-of-the-art systems using the benchmark datasets for both short and long documents, namely CNN / Daily Mail, BBC news, DUC, and arXiv are (i) abstractive text summarization using sequence to sequence RNNs by Nallapati et al [50], (ii) abstractive text summarization using neural attention by Rush et al [51], (iii) automatic extractive text summarization using autoencoders by Joshi et al [5], and (iv) abstractive text summarization using dual encoding by Yao et al [24]. The comparative analysis of ROUGE-1, ROUGE-2, and ROUGE-L scores of state-of-the-art systems with the proposed system are presented in Figure 4.



(a)



(b)



(c)

Figure. 4. Proposed System; (a) ROUGE – 1, (b) ROUGE – 2 and (c) ROUGE - L

From Figure 4, it is clear that the proposed model's ROUGE score is higher when compared to the state-of-the-art systems [13], [14], [47] and [48] in terms of Rouge-1, Rouge -2 and Rouge – L, this is because of the proposed attention free model which achieves both time efficiency and long term dependency and thus ensure long document summarization.

V. Conclusions and Future Work

In this paper, the authors proposed a new Swin-T transformer based automatic text summarization model, which involves the following stages involves data acquisition, preprocessing, Word to Vector conversion, semantic feature extraction using Swin-T transformer, clustering the similar sentences using K-medoid clustering and finally ranking the sentences and

summary generation using Bi-GRU. This setup outperforms the existing state-of-the-art systems in terms of evaluation score named ROUGE score for the short benchmark datasets as well as a long user created arXiv datasets. In the future, this work can be extended for Long multi-documents.

References

[1] Alguliev, R. M., Aliguliyev, R. M., Hajirahimova, M. S., & Mehdiyev, C. A. (2011). MCMR: Maximum coverage and minimum redundant text summarization model. Expert Systems with Applications, 38(12), 14514-14522.

[2] Moradi, M., Dorffner, G., & Samwald, M. (2020). Deep contextualized embeddings for quantifying the informative content in biomedical text summarization. Computer methods and programs in biomedicine, 184, 105117.

[3] Lloret, E. (2008). Text summarization: an overview. Paper supported by the Spanish Government under the project TEXT-MESS (TIN2006-15265-C06-01).

[4] Nenkova, A., & McKeown, K. (2012). A survey of text summarization techniques. Mining text data, 43-76.

[5] Joshi, A., Fidalgo, E., Alegre, E., & Fernández-Robles, L. (2019). SummCoder: An unsupervised framework for extractive text summarization based on deep auto-encoders. Expert Systems with Applications, 129, 200-215.

[6] Liu, L., Lu, Y., Yang, M., Qu, Q., Zhu, J., & Li, H. (2018, April). Generative adversarial network for abstractive text summarization. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

[7] Gambhir, M., & Gupta, V. (2017). Recent automatic text summarization techniques: a survey. Artificial Intelligence Review, 47, 1-66.

[8] Mishra, G., Sethi, N., & Agilandeeswari, L. (2023, March). Fuzzy Bi-GRU Based Hybrid Extractive and Abstractive Text Summarization for Long Multi-documents. In Proceedings of the 14th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2022) (pp. 153-166). Cham: Springer Nature Switzerland.

[9] Mishra, G., Sethi, N., & Agilandeeswari, L. (2023, March). Two Phase Ensemble Learning Based Extractive Summarization for Short Documents. In Proceedings of the 14th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2022) (pp. 129-142). Cham: Springer Nature Switzerland.

[10] Desai, N., & Shah, P. (2016). Automatic text summarization using supervised machine learning technique for Hindi language. Int. J. Res. Eng. Technol, 5(06), 361-367.

[11] Alguliyev, R. M., Aliguliyev, R. M., Isazade, N. R., Abdi, A., & Idris, N. (2019). COSUM: Text summarization based on clustering and optimization. Expert Systems, 36(1), e12340.

[12] Kirmani, M., Manzoor Hakak, N., Mohd, M., & Mohd, M. (2019). Hybrid text summarization: a survey. In Soft Computing: Theories and Applications: Proceedings of SoCTA 2017 (pp. 63-73). Springer Singapore.

[13] Xu, J., Gan, Z., Cheng, Y., & Liu, J. (2019). Discourse-aware neural extractive text summarization. arXiv preprint arXiv:1910.14142.

[14] Tandel, J., Mistree, K., & Shah, P. (2019, March). A review on neural network based abstractive text summarization models. In 2019 IEEE 5th International Conference for Convergence in Technology (I2CT) (pp. 1-4). IEEE.

[15] Sotudeh, S., Goharian, N., & Filice, R. W. (2020). Attend to medical ontologies: Content selection for clinical abstractive summarization. arXiv preprint arXiv:2005.00163.

[16] Su, D., Xu, Y., Yu, T., Siddique, F. B., Barezi, E. J., & Fung, P. (2020). CAiRE-COVID: A question answering and query-focused multi-document summarization system for COVID-19 scholarly information management. arXiv preprint arXiv:2005.03975.

[17] Ngamcharoen, P., Sanglerdsinlapachai, N., & Vejjanugraha, P. (2022, November). Automatic Thai Text Summarization Using Keyword-Based Abstractive Method. In 2022 17th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP) (pp. 1-5). IEEE.

[18] Cachola, I., Lo, K., Cohan, A., & Weld, D. S. (2020). TLDR: Extreme summarization of scientific documents. arXiv preprint arXiv:2004.15011.

[19] S. Wang, X. Zhao, B. Li, B. Ge and D. Tang, "Integrating Extractive and Abstractive Models for Long Text Summarization," 2017 IEEE International Congress on Big Data (BigData Congress), Honolulu, HI, USA, 2017, pp. 305-312, doi: 10.1109/BigDataCongress.2017.46.

[20] Magooda, A., & Marcjan, C. (2020). Attend to the beginning: A study on using bidirectional attention for extractive summarization. arXiv preprint arXiv:2002.03405.

[21] Sarkhel, R., Keymanesh, M., Nandi, A., & Parthasarathy, S. (2020). Interpretable multi-headed attention for abstractive summarization at controllable lengths. arXiv preprint arXiv:2002.07845.

[22] [11] Hernández-Castañeda, Ángel, et al. "Extractive automatic text summarization based on lexical-semantic keywords." IEEE Access 8 (2020): 49896-49907.

[23] [14] Yadav, Arun Kumar, et al. "Extractive text summarization using deep learning approach." International Journal of Information Technology 14.5 (2022): 2407-2415.

[24] [16] Yao, Kaichun, et al. "Dual encoding for abstractive text summarization." IEEE transactions on cybernetics 50.3 (2018): 985-996.

[25] [19] Ma, Tinghuai, et al. "T-bertsum: Topic-aware text summarization based on bert." IEEE Transactions on Computational Social Systems 9.3 (2021): 879-890.

[26] [22] Saraswathi, R. Vijaya, et al. "A LSTM based Deep Learning Model for Text Summarization."

2022 6th International Conference on Electronics, Communication and Aerospace Technology. IEEE, 2022.

[27] [28] Hark, Cengiz, and Ali Karcı. "Karcı summarization: A simple and effective approach for automatic text summarization using Karcı entropy." Information processing & management 57.3 (2020): 102187. 24

[28] [29] Belwal, Ramesh Chandra, Sawan Rai, and Atul Gupta. "Text summarization using topic-based vector space model and semantic measure." Information Processing & Management 58.3 (2021): 102536.

[29] [30] Mutlu, Begum, Ebru A. Sezer, and M. Ali Akcayol. "Candidate sentence selection for extractive text summarization." Information Processing & Management 57.6 (2020): 102359.

[30] [31]. Mohamed, Muhidin, and Mourad Oussalah. "SRL-ESA-TextSum: A text summarization approach based on semantic role labeling and explicit semantic analysis." Information Processing & Management 56.4 (2019): 1356-1372.

[31] [32]. Verma, Pradeepika, Anshul Verma, and Sukomal Pal. "An approach for extractive text summarization using fuzzy evolutionary and clustering algorithms." Applied Soft Computing 120 (2022): 108670.

[32] [33]. Mosa, Mohamed Atef, Arshad Syed Anwar, and Alaa Hamouda. "A survey of multiple types of text summarization with their satellite contents based on swarm intelligence optimization algorithms." KnowledgeBased Systems 163 (2019): 518-532.

[33] [35]. Abbasi-ghalehtaki, Razieh, Hassan Khotanlou, and Mansour Esmaeilpour. "Fuzzy evolutionary cellular learning automata model for text summarization." Swarm and Evolutionary Computation 30 (2016): 11-26.

[34] [36]. Gambhir, Mahak, and Vishal Gupta. "Recent automatic text summarization techniques: a survey." Artificial Intelligence Review 47 (2017): 1-66.

[35] [43] Alami, Nabil, et al. "Hybrid method for text summarization based on statistical and semantic treatment." Multimedia Tools and Applications 80 (2021): 19567-19600.

[36] [44]. A. Venkataramana, K. Srividya and R. Cristin, "Abstractive Text Summarization Using BART," 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India, 2022, pp. 1-6, doi: 10.1109/MysuruCon55714.2022.9972639.

[37] Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2022). Efficient transformers: A survey. ACM Computing Surveys, 55(6), 1-28.

[38] Joshi, A., Fidalgo, E., Alegre, E., & Fernández-Robles, L. (2023). DeepSumm: Exploiting topic models and sequence to sequence networks for extractive text summarization. Expert Systems with Applications, 211, 118442.

[39] Bani-Almarjeh, M., & Kurdy, M. B. (2023). Arabic abstractive text summarization using RNN-based and transformer-based architectures. Information Processing & Management, 60(2), 103227.

[40] Greene, D., & Cunningham, P. (2006, June). Practical solutions to the problem of diagonal dominance in kernel document clustering. In Proceedings of the 23rd International Conference on Machine Learning (pp. 377-384).

[41] Contractions . PyPI, 2022, https://pypi.org/project/contractions/. (Accessed 9 February 2022).

[42] Lemmatizer . spaCy API documentation, 2022, https://spacy.io/api/ lemmatizer. (Accessed 6 February 2022).

[43] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[44] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, ..., Guo B (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022

[45] S. Pandit, S. Gupta, A comparative study on distance measuring approaches for clustering, Int. J. Res. Comput. Sci. 2 (2011) 29–31, http://dx.doi.org/ 10.7815/IJORCS.21.2011.011.

[46] R. Mihalcea, P. Tarau, TextRank: Bringing order into texts | BibSonomy, in: Proceedings of EMNLP-04 and the 2004 Conference on Empirical Methods in Natural Language Processing, 2004.

[47] S.P. Lloyd, Least squares quantization in PCM, IEEE Trans. Inform. Theory 28 (1982) 129–137, http://dx.doi.org/10.1109/TIT.1982.1056489.

[48] E. Bisong, Google Colaboratory, Building machine learning and deep learning models on google cloud platform. (2019) 59–64. https://doi.org/10. 1007/978-1-4842-4470-8_7.

[49] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, 2004, pp. 74–81, https://aclanthology.org/W04-1013

[50] Nallapati, R., Zhou, B., Gulcehre, C., & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. arXiv preprint arXiv:1602.06023.

[51] Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. arXiv preprint arXiv:1509.00685.

Author Biographies

Gitanjali Mishra is a research scholar at GIET University, Gunupur, Odisha .She has completed her M.Tech from Berhampur University, Odisha in 2011.Her main areas of research interest is in NLP and Deep Learning. She has around fifteen years of teaching experience.

Nilambar Sethi received his Ph.D. in Computer Science from Berhampur University, Odisha, in 2013 and Magister degree from Utkal University, Odisha, in 2004. He is currently working as Associate Professor at the Computer Sciences Department of GIET University, Odisha.His research interests include Machine learning, Data Science, Natural language processing. In addition, he has served as a Technical Program Committee member of some international conferences and workshops. Dr. Sethi is also

an active reviewer of some renowned international journals.

Agilandeeswari L completed her Ph.D. and working as a Professor in the School of Information Technology & Engineering (SITE), VIT Vellore. She received her Bachelor's degree in Information Technology and Master's in Computer Science and Engineering from Anna University in 2005 and 2009 respectively. She is having around 20 years of teaching experience and published 70+ papers in peer-reviewed reputed journals. Her reputed publications include research articles in peer-reviewed journals namely Expert Systems with Applications, IEEE Access, Journal of Ambient Intelligence and Humanized Computing, Multimedia Tools and Applications, and Journal of Applied Remote Sensing indexing at Thomson Reuters with an average impact factor of 5. She is a peer reviewer in journals including IEEE Access, Pattern Recognition, International Journal of Remote Sensing, Array, Artificial Intelligence Review, Informatics in Medicine Unlocked, Neurocomputing, Computers, and Electrical Engineering, Journal of King Saud University–Computer and Information Sciences, IET ReView, Journal of Engineering Science and Technology (JESTEC), etc. She also published about 13 engineering books as per Anna University Syllabus. Her areas of interest include Image and video watermarking, Image processing, Neural networks, Cryptography Fuzzy Logic, Machine Learning, IoT, Information-Centric Networks, and Remote Sensing.

Yu-Chen Hu working as a Professor in the Department of Computer Science and Information Management, Providence University, Taiwan. He is having 4230 citations overall. He published 100+ papers in peer-reviewed reputed journals. He is a Editor-in-Chief and peer reviewer for a peer reviewed journals. His areas of interest include Data Compression, Image Processing, Information Hiding, Information Security and Deep Learning.