Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 07 Issue 02
2024

Research Article
https://doi.org/10.56979/702/2024

Abstractive Text Summarization for Urdu Language

Asif Raza[1*], Muhammad Hanif Soomro[2], Salahuddin[3], Inzamam Shahzad[4], and Saima Batool[5]

[1]Department of Computer Science, Bahauddin Zakariya University, Multan, Pakistan.
[2]Department of Information Technology, University of Mirpurkhas, Sindh, Pakistan.
[3]Department of Computer Science, NFC Institute of Engineering and technology, Multan, Pakistan.
[4]School of Computer Science and School of Cyberspace Science, Xiangtan University, Xiangtan, Hunan, China.
[5]Institute of Computing, Muhammad Nawaz Sharif University of Agriculture, Multan, Pakistan.
[*]Corresponding Author: Asif Raza. Email: asifraza.raza14@gmail.com

Received: March 24, 2024 **Accepted:** August 09, 2024 **Published**: September 01, 2024

Abstract: The quantity of textual data is increasing in online realm with the blink of eye and it has become very difficult to extract useful information from such enormous bundle of information. An area of natural language processing known as automated text summarization is responsible for producing gist's, abstracts, and summaries of written text in a variety of human languages. These are of a high quality and contain relevant information. Extractive and abstractive summarization are the two methods that can be used to summarize information. A lot of research is being conducted especially in extractive summary. In Urdu language, there is no research efforts in abstractive summarization up till now, so it is much needed of having research works to be done in this domain. Urdu language is mainly spoken in South Asia. In the proposed research work we use amalgam of extractive and abstractive algorithms to generate summaries. Sentence weight, TF-IDF, word frequency algorithms are used for extractive summaries. A hybrid technique is utilized so that the findings of extractive summaries can be improved. The abstractive summaries will be produced once the summaries provided by the hybrid approach have been processed using the BERT model. In order to analyses the summaries that were automatically created by the system, we have enlisted the assistance of certain Urdu language experts.

Keywords: Natural Language Processing; Term Frequency; Inverse Document Frequency; Bidirectional Encoder Representations from Transformers.

1. Introduction

After stepping into in this digital era excessive amount of text materials about any topic whether it is related to science, history, geography, sociology, psychology, sociology and so forth are easy to reachable on the Web. This world of documents, each of which stores a significant amount of information, continues to expand faster than the blink of an eye every single day. Natural Language Processing (NLP) has made great strides in recent years, and as a result, the vast majority of documents are now available in a number of different Natural Languages. The field of natural language processing (NLP) came to the forefront when people began teaching machines to translate one form of human language into another. NLP is an approach to empower computers to examine, perceive, and measure human language in a nifty and strategic manner. The incredible research work is being carried out in NLP domain. [14] [22] [50]. Most often to get some valuable and relevant information, we have to read every page of lengthy documents.

It requires a long span of time and energy to get useful and significant content. This tiresome task of reading a ton of material can mentally exhaust humans [47]. Therefore, it increases the requirement for an automatic Text summarizer because limited time is taken by text summarizers to generate summary of a text. Thus it enables user to recognize whether composed data is significant or not. In the discipline of information retrieval and NLP, Automatic text summary is the most fundamental challenge.

The method of demonstrating a ton of data or documents in a shortened form without distressing the gist of the text is text summarization. Therefore, we can say that from a set of unstructured data it is a process of finding the important text from document for the reader [19]. While making a comparison to other databases, unstructured data is available on Internet. So finding useful material itself is a hard task. The foremost goal of Text summarization is to retrieve the core subject of the data and the associate information in it and to generate summary by conserving the actual meaning of content. Due to this process we will be able to save storage space and time.

The first iterations of the summarization systems appeared in the early 1950s. The focus of the research that came before was on the word and phrase frequencies, which are considered to be the two most significant aspects of any language [1]. After that diverse algorithms of machine learning were being used for generating text summary. At present the language processing tools along with arithmetical and algebraic methods are used for generating a summary. In some of summarization systems, a manual threshold is specified to highlight the percentage of source text that will be part of the final summary.

By using text summarization systems, the main idea of lengthy documents can be judged easily and quickly. If the resultant summary has not duplicate sentences and it is emphasizing on various topics of input text, then it is considered as good summary.

Extractive summary and abstractive summary are the paradigms of text summarization system [5]. In extractive summary, the summary is generated without any alterations in the given text by extracting the important expressions from the original documents. It is like to use a highlighter to highlight key points. Normally the sequence of sentences is maintained as it was in the provided text file [37] [39]. Statistical methods like sentence length, cue based, term frequency, and scoring method etc. are used to pick sentences that are used to create summary. However, Abstractive summarizer is popular because it has the capability to highlight the key idea of text documents by generating novel sentences [49] [45]. It feels very much like writing with a pen. The challenge, however, lies in determining how to cherry-pick the essential details of a text without compromising the document's core meaning. One such issue is that computers do not yet have the capacity to understand languages as deeply as people do. This is in contrast to the human ability to understand languages. That is why, in abstractive summarization linguistic method help is taken to understand and scrutinize the original text and to generate summary [8]. Even though the both algorithms are similar as they are meant to extract key points of data. But when compared with extractive summary, the abstractive summary technique is more effective, because it creates a precise summary of text or document by itself [35]. Coherent (grammatically correct and easily readable) summary is generated by abstractive summarizer and it is the vital incentive in refining the summary’s quality.

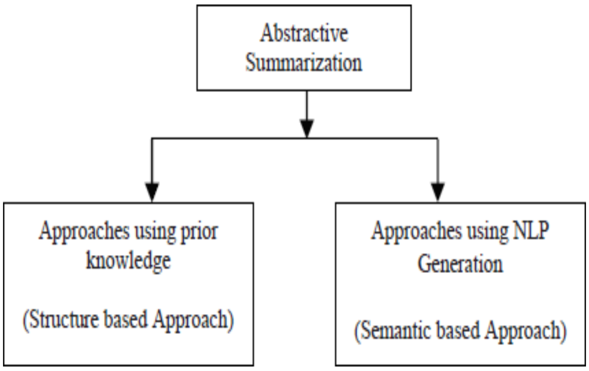
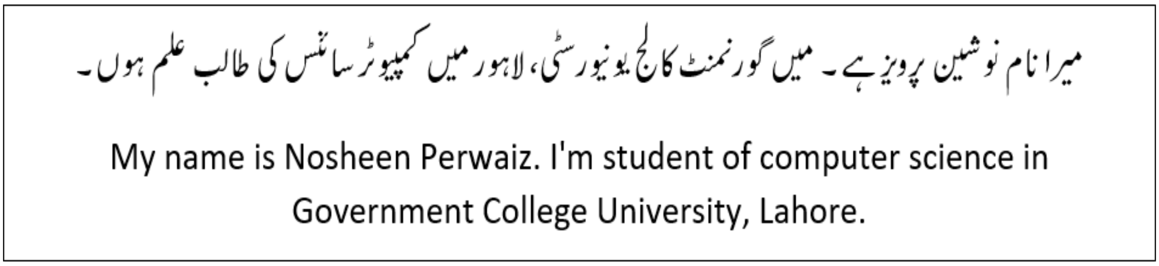

Abstractive Summarization
Approaches using prior knowledge
(Structure based Approach)
Approaches using NLP Generation
(Semantic based Approach)

Figure 1. Overview of Abstractive Summarization

Urdu - an Indo-Aryan - language that is originated from Turkish language, meaning” Lashkar” in Urdu. It is mostly spoken in South Asia. It is the national language of people of Pakistan. Urdu possesses a highly syntactic structure. There are more than **100** million Urdu speakers around the world. The lexicon of the Urdu language is heavily impacted by the Persian language as well as Sanskrit and Arabic. It possesses **38** alphabets, comprised of **26** consonants and **12** vowels in total. The indentation runs from right to left across the item. [40]. Urdu is considered as a complex language because of its morphology, diverse vocabulary and nature of words. An example of Urdu writing followed by English translation written in Pak Nastaliq font is shown in Figure 2.

At present, a lot of research work in Urdu language is carried out in the field of extractive Summarization techniques. But when compared to other languages like Arabic [27], Turkish [41], English [43], Hindi [56], Japanese [60] and Chinese [58] etc. research work done in Urdu language summaries generated by abstractive summarization approaches is in premature state. This is mainly due to lack of resources for pre-processing techniques and raw data. There is a great demand for research to be done in abstractive summarization.



میرا نام نوشین پرویز ہے۔ میں گورنمنٹ کالج یونیورسٹی، لاہور میں کمپیوٹر سائنس کی طالب علم ہوں۔

My name is Nosheen Perwaiz. I'm student of computer science in Government College University, Lahore.

Figure 2. Example of Urdu Sentence

Existing commercial and non-commercial automatic text summarizing systems for naturally occurring languages with widespread usage, such as English, French, Turkish, etc., include a large number of examples. When it comes to the Urdu language, hardly no study work in the field of artificial text summarization has been done, or the work that has been done is rather insignificant. Urdu is considered as a low resource language, because for commercial systems the resources like NLP based pre-processing tools and datasets are often not open-source or even absent. This is the main reason why it is difficult for researchers to work on abstractive summaries.

A system that uses an abstractive text summarizing technique to generate summaries of Urdu blog posts and news articles has been presented as a means of overcoming the challenges associated with Urdu text summarization. The dataset contains information that was compiled from a variety of Urdu publications, such as Express, BBC Urdu, Nawa-E-Waqt, Dawn, and Daily Jang, amongst others. The proposed technique takes advantage of the ability to replicate a condensed and more refined summary that is extremely similar to the humanoid summary. The readers of this paper will have a clear picture of the work that can be done in the field of abstractive summaries in Urdu language as a result of reading this paper. In addition to this, it will provide the opportunity and the self-assurance necessary to begin working in the Urdu language with various abstractive summarization techniques employing a wide variety of methods such as domain-based ontology, semantic graphic representation, WordNet, and many more.

Object- The prime goal of the research is to come up with a commercial framework that will generate abstractive summary of text written in Urdu language. The abstractive summary is generated after systematically consideration the content of data. This is reconstructed by compressing the input data and it looks like human generated summaries. This is implemented in three phases and two paradigms of automatic text summarization are used for this purpose. In first phase, we apply four famous techniques of extractive text summarization. In second phase we apply hybrid approach on the summaries generated in first phase. In the last phase, the abstractive summarization algorithm is applied on the summary generated in second phase. The evaluation of results is done manually by Urdu language professionals from Department of Urdu, GCU, LHR.

2. Related Work

The great majority of previous summaries of work was extractive, including the search and summary of the key phrases or sections in the source material. Summarization of text has gained a lot of awareness in latest days. We begin by discussing the associated works of extractive and abstractive summarization, followed by improved learning and self-commitment. In conclusion, this section talks about the various works that are connected to BERT (Bidirectional Encoder Representations from Transformers) and sentiment analysis.

On the other hand, humans have a tendency for retell the earliest narrative as their own language. As a consequence, human summaries are complex and ambiguous and never consist of first sentences from the text being repeated. DUC2003 and DUC-2004 competitions standardized the task of abstractive summarization.

The data for these activities is made up of news reports from a variety of subjects, one and all multiple human reference summaries. Author [15], the best performing method regarding DUC-2004 mission, applied mixture of linguistically compression techniques that are motivated and also unsupervised subject recognition algorithm which resends key-words extracted within article onto the squeeze production. Some important features of this kind include using standard machine learning Phrase-table approaches [31], weight-driven tree-transformation compression [17], and quasi-synchronous grammatical approaches [55]. Deep learning has emerged as a promising option for several NLP activities [33], and researchers have begun to view this approach as an appealing, completely data driven alternative to abstractive summarization. Author [48] encode the source with convolutional models and produce the description with context-sensitive and feed-forward neural network, yielding modern findings on the Giga-word and DUC data sets. [48] Extended this analysis by using same convolutional model for encoder as well as replacing decoder with RNN, resulting as improved functionality on both datasets. Previous studies [10] [36-38] [57] concentrated mostly on extraction methods. References [10] [36]-38] use RNN (Recurrent Neural Network) to pick sentences and obtain vector representations of sentences and papers. To compute the value of sentences, Author [57] engage RNN and GCN. Even while the extractive method can easily obtain relatively high values, the results are typically unpredictable. The sequence-to-sequence transformation is used in a number of different NLP functions, including NMT (Neural Machine Translation), QA (Question Answering), and Image Captioning (SEQ 2 SEQ) [12] [11] model was successfully used. Since the model sequence can read or make text content freely, there can be abstraction. The first work on extending the SEQ 2 SEQ attentiveness paradigm to abstraction by the researcher [49]. This strategy unquestionably works better than more traditional methods, which are shown to be inferior. The network [53] repeats terms from the source article or creates new words in the language by pointing them. References [44] have an OOV (Out-Off-Vocabulary) terms pointer network [29] in their model. Some modern approaches (e.g., [26]) cantered to SEQ 2 SEQ paradigm have been suggested, and they have all shown successful results.

While abstractive models are able to create new concepts, which makes them comparatively concise, these models also suffer from knowledge loss and have high computational costs because the input text is so extensive. The models [46] [24] suggested a hybrid architecture by combining the benefits of the extractive and abstractive methods. The extractive network extracts phrases from the input series with clear semantics, and the abstract network summarizes the extracted sentences in order to provide a final text summary.

The models that have been created on the second part of the sequence-to-sequence architecture frequently struggle with the problem of intensity bias. During the training stage, both the abstract reference and the words generated by the stage before it is utilized as the decoding input. During the test stage, however, only the words generated by the stage before it is used. Previous research [38] [9] utilized reinforcement learning [18] to address these issues. Source [38] used reinforcing training to score the phrases in pure extraction-based summarization. Reference [44] used gradient approaches to enhancement strategy learning for abstractive summarization. Reference [26] for the purpose of abstract summarization, the deep neural network method was applied. A comprehensive model of reinforced training has been created by combining a small number of human and collaborative staff members, as described in Reference [9]. Reinforcement was not employed by any of the methods, however, to bridge the non-differentiated calculation that existed between the two neural networks. In continuation of our prior efforts [46], we used strengthening learning in the model to connect the extractive and abstractive networks we had already learned.

A range of natural language operations, including machine learning [2] and summary text [53] [49], have used this target feature effectively. As we know that each element is part of the chain in a single way. Self-attention, thus, is often applied in tasks like speech processing [59], emotional analysis [30] as well as other activities. In [21], the hierarchical attention is often applied to encrypt extractive model phrases and documents. As influenced by [21], we use self-attention to represent texts.

The aforementioned model's word embedding representations are typically available in both forms: learning direct and pre-training. Learning direct obtains word representation throughout the model training phase, while pre-trained obtains term incorporation initialization by word2vec training on data set. Models [53] [46], use learning-direct, while models [24] [51] as word embedding usage words2vec. While the previous research provides a valuable analysis of the mixture of models in the system, the role of the universal language model pre-training is ignored and applied instead for the text summary model.

The pre-training model was trained using an extensive amount of unlabeled text. Model pre-training implanted vectors, be it spatial or semantic functions, are richer and more precise.

The most recent example BERTs is the model of pre-training languages, which has recently achieved success in variety like processing of natural languages activities. For the large text feedback combining word and word representations in a wide transforming, BERTs have been pre-trained [32]. Techniques for using pre-training BERTs are mostly classified as Functional and refined methods. The method is different from the literary activities [16], instead, BERTs can create better contextualized token embedding using functional based techniques, enabling our model, top of these designed, to do efficient.

Throughout this article, the term vectors of pre-training models (BERTs) are used as a job guidance, and strengthening learning is used to merge networks of extracting and producing into one model. In the first place it is needed to really understand the key concept of the material by a human-written manual definition, then pick key sentences based on the article's background details, and finally rewrite the selected sentences.

The phrase-weight system proposed by Burney, A. [7], with Word Processors' phrase weight, is completely statistical, and it is based on limiting the fact that it translated the English Stop Words (i.e., 400 Stop-words collections) and used these words to translate into Urdu and complete one of the pre-processing tasks. The central selection technique used for synthesizing text is the Cross-Language System [13] in the English-Urdu language. These summaries were translated into Urdu and used three English summary corpora. It depends on the description body of English. An algorithm was used to generate a generalized description of a single paper by an individual approach to the multi-lingual text summarizer [42]. This method uses the vector theme which divides the text and selects the higher-ranking sentence. But above all it is very effective, but the integrity of the representative consistency of the description material is not remaining, because it is not fluid at various compression ratios. Author [3] proposed the sentence fusion approach to classify repeated phrases by using the topmost local multi-sequence orientation. Sentence fusion is a multi-gene synthesis technology. In this method several documents are used as inputs, and by using the themes collection the core topic is found by processing these inputs and when the theme is finalized, the sentences are ordered and the clustering algorithm is used. The sentences are fused using sentence fusion until the sentences are ordered and a statistical synthesis is produced.

With their ideas for Chinese news synopsis, Lee et al [23] suggested fuzzy ontology, which model unknown information and thus describes field know-how accurately. Field ontology is thus defined by domain knowledge for news conferences and followed by the preliminary processing stage of this document producing significant terms in the Chinese news directory and news corpus. Fuzzy inference stage produces membership degrees for every downy concept in fuzzy field of ontology. The set of membership degrees for any flouted concept is related to different events in the field of ontology.

Tanaka et al. [52] suggested a syntactic analysis of chunks of the sentence lead and body in order to summarize newscasts. The basis of this theory is based on the techniques of phrase fusion. In order to recognize common sentences in the lead and corpse bits, the resume procedure entails inserting and substituting sentences to produce a synthesis of news broadcast by the revision of the sentence. The first step involves a syntactic analysis of the body and lead chunks accompanied trigger check pair identification, the matching sentence with various similitudes and alignment measurements. The last move consists of addition, replacement or both. The integration process includes positioning places determination, repetition control and discusses consistency review to make sure consistency and repetition removal. Replacement stage guarantees the information is increased by replacing the body sentence in the lead chunk. In order to locate semantically similar noun and verbs, Pierre-Etienne et al [4] recommended extraction rules. After extraction, the contents are selected and submit data to the generation to prevent combining candidates. It is used in a straight forward generation pattern for sentence form and vocabulary. Content-oriented summarization is done after generation.

Huong Thanh Le et al [28] suggest an approach focused on discourse law, syntactic limits, and word charts to abstract text summarization. The sentence reduction stage is dependent on input phrases, initial text keywords and syntactic restrictions. Only in a sentence mixture process, the word graph is used. In the end of every sentence and the end of every phrase, the process of generating a phrase from the important aspect. Merging of sentences is carried out through observation and adherence to a few instances. The suggested text by Ansamma Johnet al. [25] Summarizer based on the random forest classification and the characteristic ranking. The input that is pre-processed and functional ratings are then determined and the classification is trained and cross-validated, as well as the vital dimensions are summarized with the maximum marginal relevance. A conditional difficulty determining which class of the sentence is either a description or a nonreasoned class is the classification. The primary duty is the production of summary phrases from class summary. The phrases chosen based on optimum importance as well as minimal repetition.

A consensus summary is available to improve the results of the report; Ding Wang et al. [54] proposed several summary documents using a number of techniques such as the centroid process, graphics method, etc. for the assessment of various baseline combination methods such as the average ranking,

Table 1. Comparison of multi lingual developed text summarizers

Title	Language	Technique	Accuracy	Evaluation
A Neural Attention Model for Abstractive Sentence Summarization, 2015	English	Attention-Based Summarization	45%	Rouge1 Rouge2 Rouge-L
Abstractive Document Summarization via Bidirectional	Chinese	Bidirectional Decoder	40%	Rouge1 Rouge2 Rouge-L
A Text Abstraction Summary Model Based on BERT Word Embedding Learning, 2019	English	BERT model	41%	Rouge1 Rouge2 Rouge-L
Japanese Abstractive Text Summarization using BERT, 2020	Japanese	BERT mode, Pointer network	52% 53%	Rouge-N
Cross-Language Summarization System for English-Urdu	English Urdu	Extracting Sentence/ Passage	48%	Precision, recall, F1 measure
Extractive Text Summarization Models for Urdu Language	Urdu	Sentence weight, frequency	68%	Rouge1 Rouge2
Design and Development of Automatic Summarization System	Urdu	TF-IDF Lead Hybrid	58 %	Human Experts
A Rewriter Model for Urdu Document Neural Word	Urdu Embedding	Neural word	42.65%	Rouge1 Rouge2 Rouge-L

Average range, borders, median aggregation, etc. The aim is to gather data from individual summary approaches using a new weighted consensus scheme. The method for natural language generator (NLG) is fed in semantic-based technique using linguistic illustrations of documents. This approach is specifically designed to distinguish verbal phrases and verb sentences. The new graphics summarization system (Opinosis) provides portable abstract synthesis of highly redundant views [20]. It has some distinguishing features, which are essential to abstract summaries capture redundancy, subsequence gapped and systems collapsible. By scanning Opinosis graph related to subsections encoding right phrases and high points of redundancy, model produces an abstractive description. Important sentence and course score were the main components of the scheme. The path is then picked and labelled High consistency, fragmented paths and generation summary. Then two paths are declining and redundant paths are discarded.

The method incorporates the extractive and the abstractive data together to produce abstracts and that technique is based on the word graph system and compresses, merges and produces abstracts. The terminology in the paper is a sequence of vertices in the graph and the edge of the corresponding connection with two words. To establish the threshold's intensity, a page rank value weighting function was developed. The algorithm of shortest path can be applied as it gives a small sentence with more detail about the graph nodes associated with it. The key material can be found with two approaches Compendium Text Resumed approach I a group of phrases is entered as a word graph input and then sent to the image. ii) Choose essential material and apply the word graph approach from the source text.

In summary method a document or its compilation produces a succinct text that is most important key information. Single and various documents may be used as a summary task. We assume a single document or generic synthesis in the case of our Urdu text summary. Text summary methods may be classified as extractive summary and abstractive summary [6]. Extractive summary methodology derives essential phrases from text for summary production [34]. The abstract resume technique sums up the text using various vocabulary gathered from the information base based on the meaning and the semantic nature of the text. The extraction process is considered as the determination of main content in the wording, abstraction is the reformulation process, and fusion is the combination phase of extracted sections as well as the compression ensures that there is no meaningless squeezing of content [60] [61].

Based on the above analyses models that are pre-trained (BERT) word vectors which are used like task feedback, and strengthening training used to combine the extracted network and the generation network inside a prototype. First of all, we need to thoroughly grasp the principal meaning of the document, pick key phrases based on the reference details in the article, and then rewrite the chosen phrases from a human-written textbook. The concept in this paper follows the same principle and the associated relationship.

3. Methodology

Identify applicable funding agency here. If none, delete this text box.

The suggested system has its skeleton divided up into a number of different modules. In the first module, the text is pre-processed using a variety of Natural Language Processing (NLP) techniques, including normalization, tokenization, lemmatization, structured processing, and removal of stop words, amongst others. After the preliminary processing has been finished, the approaches for feature extraction are employed, and the sentences are sorted according to the weight and word/term frequencies of the individual sentences. The formation of the extractive summaries is the consequence of this action. The summaries that are produced in the second module are subjected to a hybrid method, after which a single summary is produced. The resulting summary is then submitted to the BERT model, which results in the generation of an abstractive summary. The findings of the summaries that were prepared will be evaluated in the final module. The following provides a thorough diagram of the framework that has been proposed.

3.1. Text Pre-processing

This is the first and mandatory step in any NLP tasks. There exist many open source pre-processors tools for English. For Urdu language there are also many libraries like Spacy, Inltk, Stanza and Urdu hack etc. But pre-processing with these libraries is still a difficult task. The maximum accuracy level is still not achieved.

First of all, we perform normalization of Urdu input text. For example, in Urdu there are some words in which two letters are combined together like in (,((جرأت Alif (' ا') and Hamza ('' 'أ) are separate characters but they are written in combined form. These two letters will be split. Urdu Language is really rich when it comes to its syntactic structure. For example there are some words which can be written with or without space like (، ہمت مند ہمتمند). The proper white spaces between words and punctuation will be ensured; diacritics and accents will also be removed during text normalization module.

Tokenization is the fundamental component of every text preprocessing system. Tokenization is the act of breaking down longer units of text such as sentences, phrases, paragraphs, or even entire documents into singular or individual terms. There are two primary stages involved in tokenization:

- Paragraph splitting into sentences
- Sentences splitting into words

The delimiters used for splitting of paragraphs are full stop (-), question mark (?), exclamation mark (!). However, for paragraph splitting white-spaces, quotes, commas, and semicolons are used. Tokens are created by tracing word boundaries. These word boundaries are used as starting and ending points of words. On the

Basis of these tokens, further processing like stemming and lemmatization will be performed. An example of Urdu sentence tokenization is given in the figure below.

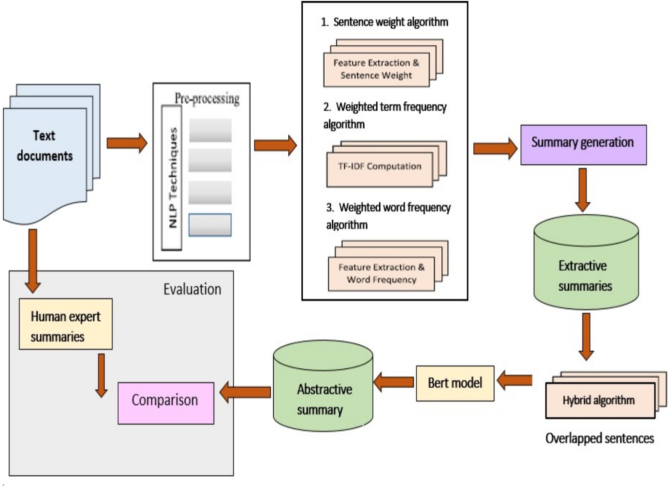
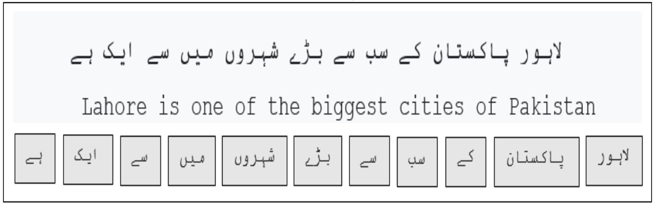


Figure 3. Diagram of Work


لاہور پاکستان کے سب سے بڑے شہروں میں سے ایک ہے

Lahore is one of the biggest cities of Pakistan

Figure 4. Sentence Tokenization

Following the completion of the tokenization process comes the subsequent step of stemming, which involves reducing the words to their most fundamental form. Take, for instance, the root word in " میرا، میری، میرے " is " "میں. The prefixes and suffixes of the word are stripped away, leaving only the root of the term. Another crucial stage in the pre-processing of text is the lemmatization of its words. It entails determining the setting in which the word is being used. Stemming and lemmatization is incomplete with POS tagging. The relative part of the speech tag is applied to tokens based on the stem and context of the word. It is a crucial step in comprehending the sentence's meaning. It is also used to extract the relationships and for building a knowledge

Graph. An example of POS tagging is given below:

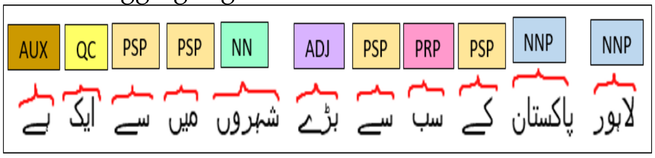
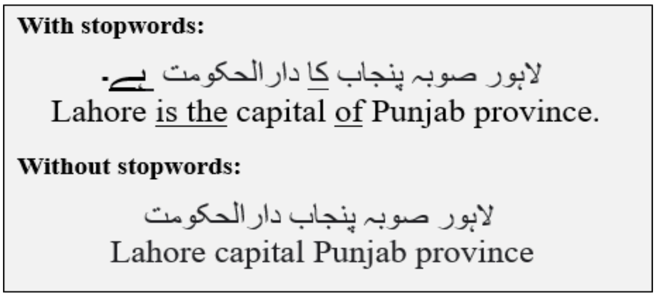


Figure 5. POS tagging

Stop-words removal is another important task in pre-processing. The most common words which occur repeatedly in all the sentences and doesn't provide any meaning to the text are called Stopwords. During the process of processing the text, the system is able to securely ignore them. The essential meaning and structure of the sentence are kept intact despite the elimination of the stopwords. Conjunctions, prepositions, and pronouns, among other types of words, might fall under this category. Some of the Urdu stopwords are کے، سے، ہےکا، کی، میں، تم،اور ، By removing stopwords from the text, the performance of the system can be improved as only meaningful and less tokens are left due to which the size of dataset and training time is decreased. A list of stopwords will be provided to the system. It will match the list with the input text and clean the data. After successfully removal of stopwords, now the remaining content words will be used for further processing. The given below figure shows the process of stopwords removal from the sentence.



With stopwords:

لاہور صوبہ پنجاب کا دارالحکومت ہے۔

Lahore is the capital of Punjab province.

Without stopwords:

لاہور صوبہ پنجاب دارالحکومت

Lahore capital Punjab province

Figure 6. Sentence with and without Stop-words

3.2. Extractive Summaries

The text is now all set for feature extraction after it has been pre-processed. Extractive summarization is the process of picking important phrases based on a set of criteria to provide a thorough summary that fully conveys the essential notion of the original text. In extractive summarizing, each sentence from the source text is assigned a weight, and only the source sentences with the highest weights are included in the final summary. There are several different approaches that can be taken to complete the task of sentence weighing. The sentence weight algorithm, the word-frequency algorithm, and the term-frequency algorithm are the three well-known methods that are utilized for the generation of extractive summaries in the system that has been proposed.

3.3. Sentence Weight Algorithm

The sentence weight algorithm is responsible for the statistical analysis that assigns a particular weight or rank to sentence in the text. The assigned weights determine whether or not the sentence is included in the final summary. Ranking of sentences is based on the amount of content words in the text divided by the number of words in the entire text. Verbs, nouns, adverbs, and adjectives are all examples of content The sentences in the input text are denoted by *T = {S1, S2, S3, S4, ..., Sn}*, n is number of sentences in entire input text T and *Sj* represents the single sentence whereas *Sw j* represents the weight of *Sj* . The next step is to compute the words count present in the entire input text. Let *W = {W1, W2, W3, w4, ...,Wn}* , n is the count of words in the sentence. The stopwords and the content words in the given text will be denoted by *Wsw* and *Wcw* respectively. The weight of the sentences will be calculated as:

$$Wcw = W - Wsw \quad (1)$$

$$Sw\ j = Wcw / Wn \times 100 \quad (2)$$

After finding the weight of all sentences, sentences will be sorted in ascending order with respect to their weight. Now we'll choose the threshold or the number of sentences we want to generate the summary. Now again sorting will be applied on the selected sentences according to the positions of sentences in the original text.

3.4. Term Frequency Algorithm

The TF-IDF statistic is a method of statistical analysis that shows the importance of a given word in relation to the input file. The significance of the TF-IDF measure rises in direct proportion to the frequency with which a certain word appears in the given text. However, it is compensated for by the amount of times a term appears in the corpus, and it helps to find that certain words are more popular than others. [Case in point] When referring to a word's overall frequency in a document, the frequency term refers to the raw frequency of the word. The inverse document frequency term also tells whether the phrase is appearing frequently or not at all in all documents. It can suggest either that it is unusual or that it appears frequently. We may calculate the overall number of phrases by dividing the number of papers that contain the term by the total number of documents.

We'll start by calculating term frequency table. Then inverse document frequency will be calculated and in the last sentence score will be calculated. Let's start from calculating term frequency matrix. It is obtained by diving the number of time a word appear in sentence with the total words in the whole document. Consider $W = \{W1, W2, W3, W4, ..., Wn\}$ or $\Sigma n\ i$=1 is the set of words in the document, where Wn is the total count of words in the given text. The count of word Wi is written as $Wj\ i$.

TF = Wji / Wn (3)

Before we can construct the IDf matrix, we need to first determine the total number of documents (Dn) and the frequency with which the Wi appears in those documents (Df). In order to normalize the frequency, we will divide (Dn) by (Df), then take the log of that number.

$IDF = log(Dn \setminus Df)$ (4)

$TF - IDF = TF \times IDF$ (5)

The significance of sentences is measured by calculating TF-IDF value of every token. The sentences are then arranged from less T-IDF values to high TF-IDF values. The sentences which have high TF-IDF value will be preferred and selected for final summary. The number of sentences included in final summary also depends on threshold selected by user. These selected sentences are arranged in the order in which they appear in input text.

3.5. Word Frequency Summarization

Only the frequency of the terms found inside the content will be calculated using this method. In the first step of this process, we will compute the frequency table by first removing any stop words and then only adding the content words from the document. Let Wf equal the number of content words in a sentence (W1, W2, W3,...,Wn), where Wfi is the frequency of the word Wi and Wn is the total number of words in a phrase. Add up the number of times each word appears in the sentence, then divide that total by the total number of words in the sentence. This will give you the sentence score, or SC.

$SC = Wfi / Wn$ (6)

The possible problem with this algorithm is the long phrases benefit over short phrases. We split each sentence score by the number of terms in the sentence to solve this problem.

WF = SC / Wn (7)

3.6. Hybrid Algorithm

We used the three summaries of the input document generated by sentence weight algorithm, TF-IDF algorithm and word frequency algorithm to compose hybrid summary of the document. The common sentences present in the summaries generated by three algorithms will be chosen to generate single summary. This shows the importance of sentences as they occur in all three summaries. This approach is used to refine the results of extractive summaries. The precise and articulate summary will be generated from long text.

In order to generate hybrid summary, a threshold between 30 and 40 percent of the input text is supplied. If the length of the summary is lower than the threshold after taking into account all of the common sentences extracted from all three algorithms, then it will select the common sentences from the two extractive summaries. We are going to create sets of indexes for the sentences. We are going to keep a record of the index number for each sentence. Let's say A = s1, s2, s3,...., sn, where each of these numbers represents an index of a sentence found in the given material. All three algorithms will do a comparison of these indices with the sentences that are included in the summaries that they generate.

Let SW is the set of indexes of sentences included in the summary generated by sentence weight algorithm. $SW = \{\{s'1\}, \{s'2\}, \{s'3\}, \ldots\ldots\{s'n\}\}$. The set of indexes of sentences included in summary generated by $TF–IDF = \{\{s''1\}, \{s''2\}, \{s''3\}, \ldots\ldots\{s''n\}\}$.

We'll compare the indexes and the common indexes will be stored in FS.

HS = $SW \cap TF - IDF$ (8)

The indexes of sentences by word frequency algorithm will be stored in $WF = \{\{s'''1\}, \{s'''2\}, \{s'''3\}, \ldots\ldots\{s'''n\}\}$. We'll compare FS and WF. The sentences whose indexes will match will be taken for final summary.

HS = $SW \cap TF - IDF \cap WF$ (9)

3.7. Abstractive Summary

The algorithms that are used to generate an abstractive summary do not collect sentences from the source material; rather, they aim to capture the essential concepts of the text and construct new phrases to express them. The process of abstractive summarization is quite similar to the way that human summarizers approach their work. The proposed system summarized the information supplied by the

Hybrid algorithm by employing a model known as BERT, which stands for "Bidirectional Encoder Representations from Transformers." These models are often regarded as the most effective method currently available for carrying out NLP-related activities. Since BERT models have already been trained on enormous data-sets, there is no need for any extra training to be performed. In order to deliver the highest quality summaries, it makes use of a sturdy flat design that incorporates transition layers between sentences.

It uses a typical seq2seq architecture including bidirectional encoder, decoder, embedding and summarization layers. The overview architecture of BERT model is given:

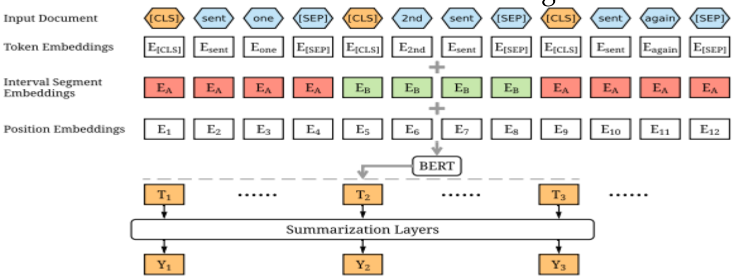


Figure 7. Overview architecture of Bert Model

Tem is competent as a masked model. In contrast to other extractive summarizers, it uses embedding for various phrases. These embedding are adjusted to create the relevant summaries. in proposed work the labels X and S are passed to BERT instead of many phrases. The source document is represented as *X* = *{x*1, *x*2, ..., *xn}*. The sequence of sentences form source material is denoted as *S* = *{s*1, *s*2, ..., *sn}*. Let *yi* = *{*0, 1*}* are the two possibilities which signifies whether a certain sentence will be selected or not.

3.8. Encoder

The sentences of input material are encoded in order to be preprocessed. An LSTM encoder is used in proposed work. We'll use two tags; a CLS tag before each phrase and followed by SEP tag. CLS tag is responsible for aggregating the features of sentences. Urdu Input tokens are passes to encoder one by one and it stores the summarized data in hidden layer. We'll only save the data of internal layer and output will be discarded. This layer encapsulates the data of all input sequences so that the accurate predictions can be made by decoder. The formula used to calculate hidden layer is:

$$ht = f(W(hh)ht{-}1 + W(hx)xt) \quad (10)$$

3.9. Decoder

Decoder is also an LSTM model. We'll use the final state of encoder as initial state of decoder. The decoder begins to generate output sequence by using these states and the outputs will be stored for further use. We'll use the hidden states of encoder to generate output of decoder and its own hidden state. In the hidden state of decoder word embedding are fed at t-th time and can be calculated as:

$$ht = f(W(hh)ht{-}1 \quad (11)$$

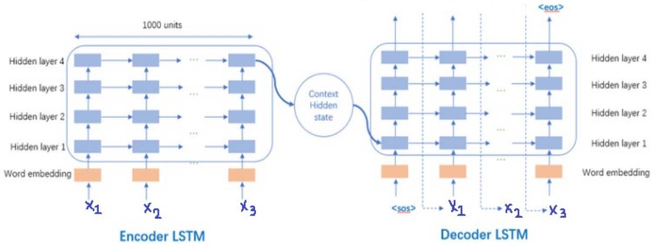


Figure. 8. Encoder-Decoder in LSTM

Table 2. Title and Type

Sr. No	Title of Document	Type
1	کورونا کی خراب ہوتی صورتحال	Current Affairs
2	کلائمیٹ چینج یا موسمیاتی تبدیلیاں	Environment
3	یوم مئی ، مزدور کا عالمی دن یا لیبر ڈے	Social issues

4	صحت پر ماحول کا اثر	Health
5	ہم، اسلام اور پاکستان	Religion
6	الیکشن سے پہلےسلیکشن	Politics
7	7ویں ایشیائی سرما کھیلوں کا اختتام	Sports
8	وادی چترال	Tourism
9	مائیکرو کمپیوٹرز کی دنیا	Technology
10	اصحابِ کہف	History

3.10. The Model

The sequence of sentences S = fs (X) mod 2 is computed as fs (x) is the total count of sentences in x. The resulting BERT encoder output is denoted as M. Next step is to input M and the decoder's output at the t-th time stage. As demonstrated in (1), the possibility of the vocabulary can be calculated at the t-th time stage. The loss in training L can be calculated using (1).

$$Pt(w) = fdec\ (w|M, X<t) \tag{12}$$

$$L = -\Sigma ni=1\ logP\ (Xt\ |M,Ht-1) \tag{13}$$

4. Result & Decision

Assessment of results for summaries generated in Urdu language is the most difficult task as standard datasets are not available. However, an Urdu summary corpus is available for abstractive summaries which is comparatively small. These summaries are produced without any standard guidelines. The average size ratio of resultant summary is between 33 and 40%. In the discussed corpus, some summaries have size up to 80% of input material. There are fifty articles divided in many categories like health, news, history, religion, current affairs, technology, sports and tourism etc. But the length of various articles is less than 400-500 words which is not recommended to generate abstractive summaries. With these

4.1. System generated Results

Verifying that the summaries that were generated by the machine are accurate is an exceedingly challenging task. On the other hand, the compression rate of system generated summaries can be determined with the use of certain computational formulas. It provides an estimate of the total number of sentences that are included in the ensuing summary in relation to the length of the original text.

Compression-rate% = total-words-summary ×100 / input-text-length

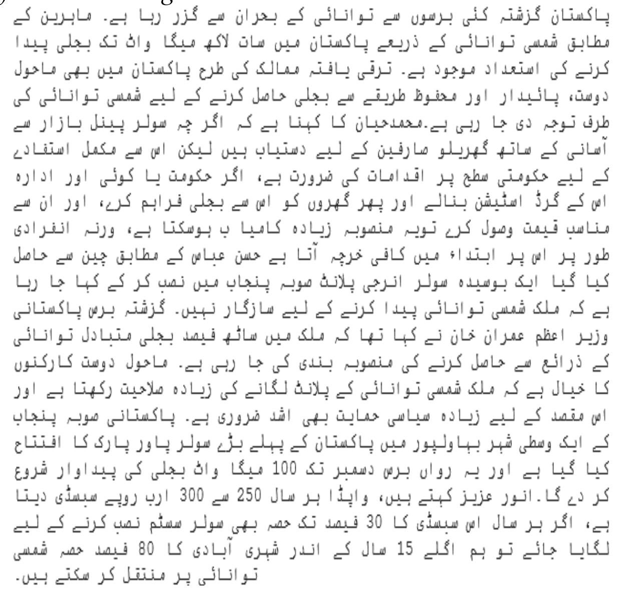
Some tests are conducted by submitting some articles to the text summarizer. These experiments are performed on our collected dataset of 50 articles which are also divided into the same categories as discussed in the referenced corpus. The articles are collected from BBC Urdu, Express Urdu Blogs, Jang news, Nawae-Waqt and some other Urdu magazines. The word count for these articles ranges anywhere from 800 to 1600 total characters. The size ratio of the extracted summary is almost identical to that of the input text. The abstractive model is then given these data after they have been compressed so that it can generate a summary. This data is compressed once more, and the size of the resulting summary is almost 20% of the size of the given text. The resultant summaries are easy to read, and the summary's linguistic integrity is preserved. The key concept of the input material is presented solely in the approximate description. It displayed precise data while retaining the original text's content.

4.2. Evaluation

To assess the other features like paraphrasing, main idea, punctuation, grammar and accuracy (sentences from input material) of generated summaries, services from three Urdu Experts are taken. These people are experienced and having knowledge in the domain of Urdu Language. The author provided them the dataset and the system generated summaries and requested them to assess the resultant summaries according to these metrics. It will give a rough idea of human's perspective of creating summaries and the characteristics on the basis of which they create summaries. Even though it is a tremendous challenge for a computer to 100% comply with human language, their mother 's tongue skills and vocabulary.

4.3. Sample text of single document articles and their Summaries

4.3.1. Summary Generated by Sentence Weight

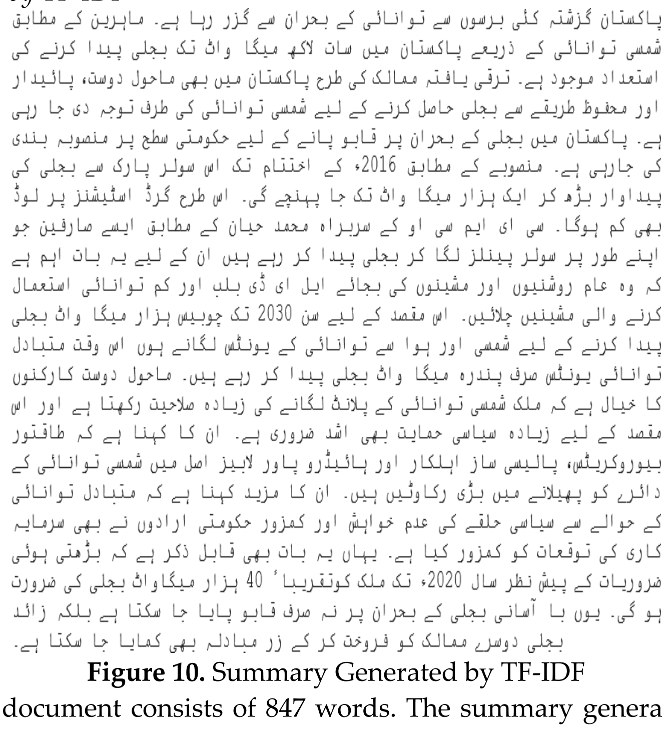


پاکستان گزشتہ کئی برسوں سے توانائی کے بحران سے گزر رہا ہے۔ ماہرین کے
مطابق شمسی توانائی کے ذریعے پاکستان میں سات لاکھ میگا واٹ تک بجلی پیدا
کرنے کی استعداد موجود ہے۔ ترقی یافتہ ممالک کی طرح پاکستان میں بھی ماحول
دوست، پائیدار اور محفوظ طریقے سے بجلی حاصل کرنے کے لیے شمسی توانائی کی
طرف توجہ دی جا رہی ہے۔محمدحیان کا کہنا ہے کہ اگر چہ سولر پینل بازار سے
آسانی کے ساتھ گھریلو صارفین کے لیے دستیاب ہیں لیکن اس سے مکمل استفادے
کے لیے حکومتی سطح پر اقدامات کی ضرورت ہے، اگر حکومت یا کوئی اور ادارہ
اس کے گرڈ اسٹیشن بنالے اور پھر گھروں کو اس سے بجلی فراہم کرے، اور ان سے
مناسب قیمت وصول کرے تویہ منصوبہ زیادہ کامیا ب ہوسکتا ہے، ورنہ انفرادی
طور پر اس پر ابتداء میں کافی خرچہ آتا ہے حسن عباس کے مطابق چین سے حاصل
کیا گیا ایک بوسیدہ سولر انرجی پلانٹ صوبہ پنجاب میں نصب کر کے کہا جا رہا
ہے کہ ملک شمسی توانائی پیدا کرنے کے لیے سازگار نہیں۔ گزشتہ برس پاکستانی
وزیر اعظم عمران خان نے کہا تھا کہ ملک میں ساٹھ فیصد بجلی متبادل توانائی
کے ذرائع سے حاصل کرنے کی منصوبہ بندی کی جا رہی ہے۔ ماحول دوست کارکنوں
کا خیال ہے کہ ملک شمسی توانائی کے پلانٹ لگانے کی زیادہ صلاحیت رکھتا ہے اور
اس مقصد کے لیے زیادہ سیاسی حمایت بھی اشد ضروری ہے۔ پاکستانی صوبہ پنجاب
کے ایک وسطی شہر بہاولپور میں پاکستان کے پہلے بڑے سولر پاور پارک کا افتتاح
کیا گیا ہے اور یہ رواں برس دسمبر تک 100 میگا واٹ بجلی کی پیداوار شروع
کر دے گا۔انور عزیز کہتے ہیں، واپڈا ہر سال 250 سے 300 ارب روپے سبسڈی دیتا
ہے، اگر ہر سال اس سبسڈی کا 30 فیصد تک حصہ بھی سولر سسٹم نصب کرنے کے لیے
لگایا جائے تو ہم اگلے 15 سال کے اندر شہری آبادی کا 80 فیصد حصہ شمسی
توانائی پر منتقل کر سکتے ہیں۔

Figure. 9. Summary Generated by Sentence Weight

The sample input document consists of 847 words. The summary generated by sentence weight algorithm consists of 325 words which is almost 38% of input text. The sentences with highest weight are picked and arranged in ascending order. The main idea of input document is preserved in this summary

4.3.2. Summary Generated by TF-IDF



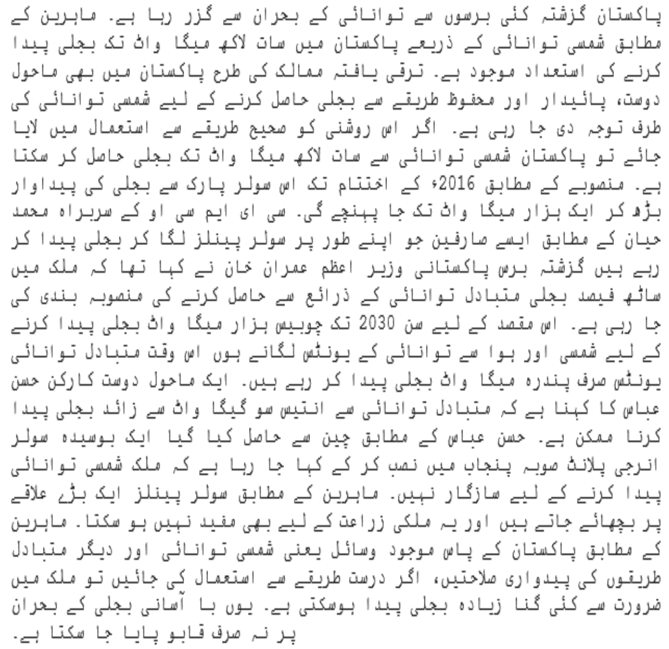
پاکستان گزشتہ کئی برسوں سے توانائی کے بحران سے گزر رہا ہے۔ ماہرین کے مطابق
شمسی توانائی کے ذریعے پاکستان میں سات لاکھ میگا واٹ تک بجلی پیدا کرنے کی
استعداد موجود ہے۔ ترقی یافتہ ممالک کی طرح پاکستان میں بھی ماحول دوست، پائیدار
اور محفوظ طریقے سے بجلی حاصل کرنے کے لیے شمسی توانائی کی طرف توجہ دی جا رہی
ہے۔ پاکستان میں بجلی کے بحران پر قابو پانے کے لیے حکومتی سطح پر منصوبہ بندی
کی جارہی ہے۔ منصوبے کے مطابق 2016ء کے اختتام تک اس سولر پارک سے بجلی کی
پیداوار بڑھ کر ایک ہزار میگا واٹ تک جا پہنچے گی۔ اس طرح گرڈ اسٹیشنز پر لوڈ
بھی کم ہوگا۔ سی ای ایم سی او کے سربراہ محمد حیان کے مطابق ایسے صارفین جو
اپنے طور پر سولر پینلز لگا کر بجلی پیدا کر رہے ہیں ان کے لیے یہ بات اہم ہے
کہ وہ عام روشنیوں اور مشینوں کی بجائے ایل ای ڈی بلب اور کم توانائی استعمال
کرنے والی مشینیں چلائیں۔ اس مقصد کے لیے سن 2030 تک چوبیس ہزار میگا واٹ بجلی
پیدا کرنے کے لیے شمسی اور ہوا سے توانائی کے یونٹس لگانے ہوں اس وقت متبادل
توانائی یونٹس صرف پندرہ میگا واٹ بجلی پیدا کر رہے ہیں۔ ماحول دوست کارکنوں
کا خیال ہے کہ ملک شمسی توانائی کے پلانٹ لگانے کی زیادہ صلاحیت رکھتا ہے اور اس
مقصد کے لیے زیادہ سیاسی حمایت بھی اشد ضروری ہے۔ ان کا کہنا ہے کہ طاقتور
بیوروکریٹس، پالیسی ساز اہلکار اور ہائیڈرو پاور لابیز اصل میں شمسی توانائی کے
دائرے کو پھیلانے میں بڑی رکاوٹیں ہیں۔ ان کا مزید کہنا ہے کہ متبادل توانائی
کے حوالے سے سیاسی حلقے کی عدم خواہش اور کمزور حکومتی ارادوں نے بھی سرمایہ
کاری کی توقعات کو کمزور کیا ہے۔ یہاں یہ بات بھی قابل ذکر ہے کہ بڑھتی ہوئی
ضروریات کے پیش نظر سال 2020ء تک ملک کوتقریباً 40 ہزار میگاواٹ بجلی کی ضرورت
ہو گی۔ یوں با آسانی بجلی کے بحران پر نہ صرف قابو پایا جا سکتا ہے بلکہ زائد
بجلی دوسرے ممالک کو فروخت کر کے زر مبادلہ بھی کمایا جا سکتا ہے۔

Figure 10. Summary Generated by TF-IDF

The sample input document consists of 847 words. The summary generated by TF-IDF algorithm consists of 348 words which is almost 41% of input text. The sentences in summary are exactly taken from the source material and they present.

4.3.3. *Summary Generated by Word Frequency*

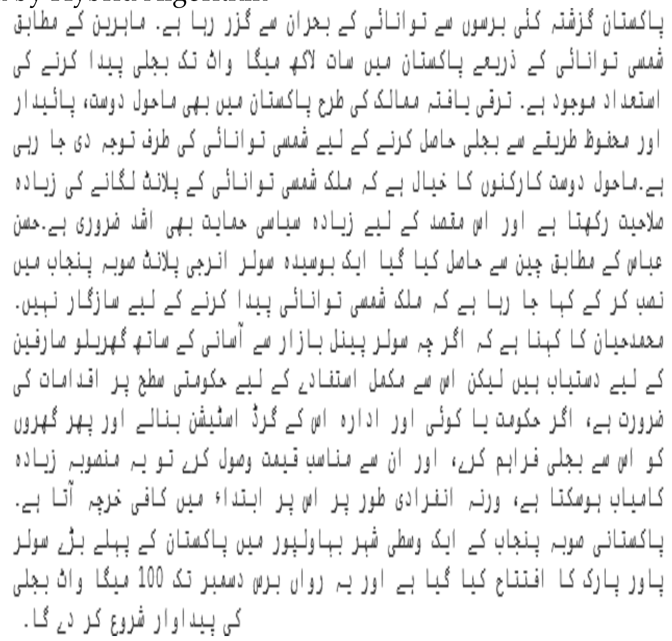
The sample input document consists of 847 words. The summary generated by TF-IDF algorithm consists of 339 words which is almost 40% of input text. The sentences with word which are most repeated will be picked. Sentences in summary are exactly taken from the source material and depicts the key concept of input file.



پاکستان گزشتہ کئی برسوں سے توانائی کے بحران سے گزر رہا ہے۔ ماہرین کے
مطابق شمسی توانائی کے ذریعے پاکستان میں سات لاکھ میگا واٹ تک بجلی پیدا
کرنے کی استعداد موجود ہے۔ ترقی یافتہ ممالک کی طرح پاکستان میں بھی ماحول
دوست، پائیدار اور محفوظ طریقے سے بجلی حاصل کرنے کے لیے شمسی توانائی کی
طرف توجہ دی جا رہی ہے۔ اگر اس روشنی کو صحیح طریقے سے استعمال میں لایا
جائے تو پاکستان شمسی توانائی سے سات لاکھ میگا واٹ تک بجلی حاصل کر سکتا
ہے۔ منصوبے کے مطابق 2016ء کے اختتام تک اس سولر پارک سے بجلی کی پیداوار
بڑھ کر ایک ہزار میگا واٹ تک جا پہنچے گی۔ سی ای ایم سی او کے سربراہ محمد
حیان کے مطابق ایسے صارفین جو اپنے طور پر سولر پینلز لگا کر بجلی پیدا کر
رہے ہیں گزشتہ برس پاکستانی وزیر اعظم عمران خان نے کہا تھا کہ ملک میں
ساٹھ فیصد بجلی متبادل توانائی کے ذرائع سے حاصل کرنے کی منصوبہ بندی کی
جا رہی ہے۔ اس مقصد کے لیے سن 2030 تک چوبیس ہزار میگا واٹ بجلی پیدا کرنے
کے لیے شمسی اور ہوا سے توانائی کے یونٹس لگانے ہوں اس وقت متبادل توانائی
یونٹس صرف پندرہ میگا واٹ بجلی پیدا کر رہے ہیں۔ ایک ماحول دوست کارکن حسن
عباس کا کہنا ہے کہ متبادل توانائی سے انتیس سو گیگا واٹ سے زائد بجلی پیدا
کرنا ممکن ہے۔ حسن عباس کے مطابق چین سے حاصل کیا گیا ایک بوسیدہ سولر
انرجی پلانٹ صوبہ پنجاب میں نصب کر کے کہا جا رہا ہے کہ ملک شمسی توانائی
پیدا کرنے کے لیے سازگار نہیں۔ ماہرین کے مطابق سولر پینلز ایک بڑے علاقے
پر بچھائے جاتے ہیں اور یہ ملکی زراعت کے لیے بھی مفید نہیں ہو سکتا۔ ماہرین
کے مطابق پاکستان کے پاس موجود وسائل یعنی شمسی توانائی اور دیگر متبادل
طریقوں کی پیدواری صلاحتیں، اگر درست طریقے سے استعمال کی جائیں تو ملک میں
ضرورت سے کئی گنا زیادہ بجلی پیدا ہوسکتی ہے۔ یوں با آسانی بجلی کے بحران
پر نہ صرف قابو پایا جا سکتا ہے۔

Figure 11. Summary Generated by Word frequency

4.4. Summary Generated by Hybrid Algorithm



پاکستان گزشتہ کئی برسوں سے توانائی کے بحران سے گزر رہا ہے۔ ماہرین کے مطابق
شمسی توانائی کے ذریعے پاکستان میں سات لاکھ میگا واٹ تک بجلی پیدا کرنے کی
استعداد موجود ہے۔ ترقی یافتہ ممالک کی طرح پاکستان میں بھی ماحول دوست، پائیدار
اور محفوظ طریقے سے بجلی حاصل کرنے کے لیے شمسی توانائی کی طرف توجہ دی جا رہی
ہے۔ماحول دوست کارکنوں کا خیال ہے کہ ملک شمسی توانائی کے پلانٹ لگانے کی زیادہ
صلاحیت رکھتا ہے اور اس مقصد کے لیے زیادہ سیاسی حمایت بھی اشد ضروری ہے۔حسن
عباس کے مطابق چین سے حاصل کیا گیا ایک بوسیدہ سولر انرجی پلانٹ صوبہ پنجاب میں
نصب کر کے کہا جا رہا ہے کہ ملک شمسی توانائی پیدا کرنے کے لیے سازگار نہیں۔
محمدحیان کا کہنا ہے کہ اگر چہ سولر پینل بازار سے آسانی کے ساتھ گھریلو صارفین
کے لیے دستیاب ہیں لیکن اس سے مکمل استفادے کے لیے حکومتی سطح پر اقدامات کی
ضرورت ہے، اگر حکومت یا کوئی اور ادارہ اس کے گرڈ اسٹیشن بنالے اور پھر گھروں
کو اس سے بجلی فراہم کرے، اور ان سے مناسب قیمت وصول کرے تو یہ منصوبہ زیادہ
کامیاب ہوسکتا ہے، ورنہ انفرادی طور پر اس پر ابتداء میں کافی خرچہ آتا ہے۔
پاکستانی صوبہ پنجاب کے ایک وسطی شہر بہاولپور میں پاکستان کے پہلے بڑے سولر
پاور پارک کا افتتاح کیا گیا ہے اور یہ رواں برس دسمبر تک 100 میگا واٹ بجلی
کی پیداوار شروع کر دے گا۔

Figure 12. Hybrid Summary

Here hybrid summary consists of 241 words which is almost one-third of the given text. The common sentences of three summaries are picked and a refined summary is generated which only consists of the most important sentences of input document.

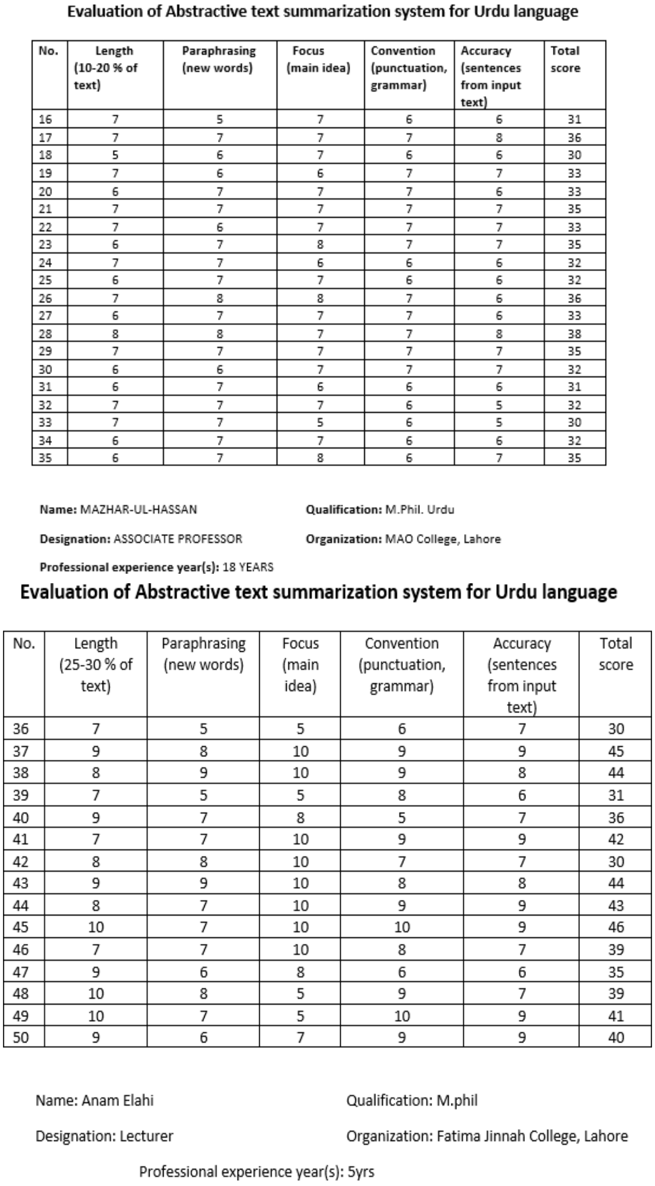
4.5. Abstractive Summary

The generated summary is almost one-fourth of the hybrid summary. Because Hybrid summary is provided to BERT model for abstractive summary generation. There are many new words we can see in the summary. The focus, convention and accuracy is also maintained. We can say that the generated summary is compact and concise. Now there is no need to read the data of 2 pages to know what is written inside it. You can just read 4-5 lines to find whether the document is useful for you or not.

4.6. Evaluation Results

The metrics for evaluation used in the existing models; Rouge1, Rouge2, Rouge-L are based on precision, recall and F1 measure for finding the accuracy. In these models the human and system generated summaries were being compared by the system but that was not the finest one. In the present model the accuracy is measured by generating the summaries by the system and assessing the results by the humans. This method is better than the previous because the understanding of a language in terms of tongue skills and vocabulary of the humans with 'Urdu' mother language is finer than the machines.

پاکستان ماحول دوست ، پائیدار اور محفوظ طر یقے سے بجلی پیدا کرنے کے لئے شمسی توانائی پر توجہ مرکوز کررہا ہے۔ طر یقوں کی
پیداواری گنجائش ، اگر مناسب طر یقے سے استعمال کی جائے تو ، ملک کی ضرورت سے ز یادہ بجلی پیدا ہوسکتی ہےپاکستان شمسی توانائی
کے ذر یعہ 000،700 میگا واٹ تک بجلی پیدا کرنے کی صلاحیت رکھتا ہےپاکستان کے صوبہ پنجاب کے وسطی شہر بہاولپور میں پاکستان کے
پہلے بڑے شمسی توانائی پارک کا افتتاح کیا گیا ہے



Evaluation of Abstractive text summarization system for Urdu language

No.	Length (10-20 % of text)	Paraphrasing (new words)	Focus (main idea)	Convention (punctuation, grammar)	Accuracy (sentences from input text)	Total score
16	7	5	7	6	6	31
17	7	7	7	7	8	36
18	5	6	7	6	6	30
19	7	6	6	7	7	33
20	6	7	7	7	6	33
21	7	7	7	7	7	35
22	7	6	7	7	7	33
23	6	7	8	7	7	35
24	7	7	6	6	6	32
25	6	7	7	6	6	32
26	7	8	8	7	6	36
27	6	7	7	7	6	33
28	8	8	7	7	8	38
29	7	7	7	7	7	35
30	6	6	7	7	7	32
31	6	7	6	6	6	31
32	7	7	7	6	5	32
33	7	7	5	6	5	30
34	6	7	7	6	6	32
35	6	7	8	6	7	35

Name: MAZHAR-UL-HASSAN **Qualification:** M.Phil. Urdu

Designation: ASSOCIATE PROFESSOR **Organization:** MAO College, Lahore

Professional experience year(s): 18 YEARS

Evaluation of Abstractive text summarization system for Urdu language

No.	Length (25-30 % of text)	Paraphrasing (new words)	Focus (main idea)	Convention (punctuation, grammar)	Accuracy (sentences from input text)	Total score
36	7	5	5	6	7	30
37	9	8	10	9	9	45
38	8	9	10	9	8	44
39	7	5	5	8	6	31
40	9	7	8	5	7	36
41	7	7	10	9	9	42
42	8	8	10	7	7	30
43	9	9	10	8	8	44
44	8	7	10	9	9	43
45	10	7	10	10	9	46
46	7	7	10	8	7	39
47	9	6	8	6	6	35
48	10	8	5	9	7	39
49	10	7	5	10	9	41
50	9	6	7	9	9	40

Name: Anam Elahi Qualification: M.phil

Designation: Lecturer Organization: Fatima Jinnah College, Lahore

Professional experience year(s): 5yrs

Figure 13. Abstractive Summary

5. Conclusion and Future Work

A plethora of information text summarizers consciously participate in understanding only the key idea of source article or other document in any language. People typically tend to read the highlights of news articles, movie premieres, or an overview of latest developments in science journals, etc. from the gist content found on various web pages or other online portals. Because of the abundance of online information today, experts in Natural Language Processing have focused on meeting the need for automated summarization.

Even though Urdu is national language of Pakistan. It is globally written and spoken by billions of people. Still its resources are low with regards to language processing research. There are a lot of web portals or news websites which are generating Urdu data on daily basis such as Express news. In order to save time, people choose to read abstract or synopsis of lengthy data. For English language summary, there are several approaches available. Conversely, for Urdu language these strategies do not exists publically.

In this research, first the data is cleaned by pre-processing techniques. Then important sentences are extracted by applying sentence weight, TF-IDF, and word frequency algorithms. Sentence rank method is used to assess the summaries generated by these algorithms. After the extractive summaries are generated, a hybrid algorithm is applied to refine the results. The common sentences from the summaries generated by famous extractive algorithms are picked and a single unit summary is produced which represents the main idea of source material. The hybrid summary is then processed by the pre-trained BERT model to generate abstractive summary. Experiments are executed on author's own collected dataset of 50 articles divided into different categories like health, news, sports etc. Evaluation is done by Urdu professionals in order to find accuracy of proposed system.

In this proposed work, summaries are generated for single document. We'll try to improve our work by generating multi-document summaries in future. Moreover, the window-based application forms will be introduced to make this proposed system commercial. We'll try to add more vocabulary and synonyms to refine the results and to make summaries closer to human generated summaries.

References

1. Alaa F. Alsaqer and S. Sasi. Movie review summarization and sentiment analysis using rapidminer. 2017 International Conference on Networks and Advances in Computational Technologies, pages 329–335, 2017.
2. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.
3. Regina Barzilay and Kathleen R McKeown. Sentence fusion for multidocument news summarization. Computational Linguistics, 31(3):297–328, 2005
4. Riadh Belkebir and Ahmed Guessoum. A supervised approach to arabic text summarization using adaboost. In New contributions in information systems and technologies, pages 227–236. Springer, 2015.
5. Neelima Bhatia and Arunima Jaiswal. Article: Trends in extractive and abstractive techniques in text summarization. International Journal of Computer Applications, 117(6):21–24, May 2015.
6. Aqil Burney, Badar Sami, Nadeem Mahmood, Zain Abbas, and Kashif Rizwan. Urdu text summarizer using sentence weight algorithm for word processors. International Journal of Computer Applications, 46(19):38–43, 2012.
7. A. Jaya C. Sunitha and Amal Ganesh. A study on abstractive summarization techniques in indian languages. Procedia Computer Science, 87:25–31, 2016. ISSN 1877-0509.
8. Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. Deep communicating agents for abstractive summarization. arXiv preprint arXiv:1803.10357, 2018.
9. Jianpeng Cheng and Mirella Lapata. Neural summarization by extracting sentences and words. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 484–494, Berlin, Germany, aug 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1046. URL https://www.aclweb.org/anthology/ P16-1046.
10. Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078, 2014
11. Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 895–903, 2017.
12. Amna Yasin CIIT. Cross-Language Summarization System for English-Urdu Language. PhD thesis, 2016.
13. Dipanjan Das and André Martins. A survey on automatic text summarization. 12 2007.
14. Hal Daumé III and Daniel Marcu. Induction of word and phrase alignments for automatic document summarization. Computational Linguistics, 31(4): 505–530, 2005.
15. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
16. Sevgi Dogan, David Arditi, and Hüsnü Günaydın. Using decision trees for determining attribute weights in a case-based model of early cost prediction. Journal of Construction Engineering and Management-asce - J CONSTR ENG MANAGE-ASCE, 134, 02 2008. doi: 10.1061/(ASCE)0733-9364(2008) 134:2(146).
17. Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G Bellemare, and Joelle Pineau. An introduction to deep reinforcement learning. arXiv preprint arXiv:1811.12560, 2018.
18. Gupta Vishal Gambhir, Mahak. Recent automatic text summarization techniques: a survey, volume 47. Springer, 2017.
19. Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. 2010.
20. Paul Gastin and Denis Oddoux. Fast ltl to büchi automata translation. In Computer Aided Verification, pages 53–65. Springer, 2001.
21. Md. Majharul Haque, Suraiya Pervin, and Zerina Begum. Literature review of automatic multiple documents text summarization. International Journal of Innovation and Applied Studies, 3:121–129, 05 2013.
22. Pedro Hípola, José A Senso, Amed Leiva-Mederos, and Sandor DomínguezVelasco. Ontology-based text summarization. the case of texminer. Library hi tech, 2014.
23. Younes Jaafar and Karim Bouzoubaa. Towards a new hybrid approach for abstractive summarization. Procedia computer science, 142:286–293, 2018.
24. Ansamma John and M Wilscy. Random forest classifier based multi-document summarization system. In 2013 IEEE Recent Advances in Intelligent Computational Systems (RAICS), pages 31–36. IEEE, 2013.

25. Enise Karakoç and Burcu Yılmaz. Deep learning based abstractive turkish news summarization. In 2019 27th Signal Processing and Communications Applications Conference (SIU), pages 1–4. IEEE, 2019.
26. Samira Lagrini, Mohammed Redjimi, and Nabiha Azizi. Automatic arabic text summarization approaches. International Journal of Computer Applications, 164(5):31–37, Apr 2017. ISSN 0975-8887. doi: 10.5120/ ijca2017913628. URL http://www.ijcaonline.org/archives/volume164/ number5/27480-2017913628.
27. Huong Thanh Le and Tien Manh Le. An approach to abstractive text summarization. In 2013 International Conference on Soft Computing and Pattern Recognition (SoCPaR), pages 371–376. IEEE, 2013.
28. Jian Li, Yue Wang, Michael R Lyu, and Irwin King. Code completion with neural attention and pointer networks. arXiv preprint arXiv:1711.09573, 2017.
29. Zheng Li, Ying Wei, Yu Zhang, and Qiang Yang. Hierarchical attention transfer network for cross-domain sentiment classification. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.
30. Lidia S. Chao Francisco Oliveira Liang Tian, Derek F. Wong. A relationship: Word alignment, phrase table, and translation quality. In The Scientific World Journal, page 13, 2014.
31. Yang Liu. Fine-tune BERT for extractive summarization. CoRR, abs/1903.10318, 2019. URL http://arxiv.org/abs/1903.10318.
32. Asif, Sohaib, et al. "MozzieNet: A deep learning approach to efficiently detect malaria parasites in blood smear images." International Journal of Imaging Systems and Technology 34.1 (2024): e22953.
33. N Moratanch and S Chitrakala. A survey on extractive text summarization. In 2017 international conference on computer, communication and signal processing (ICCCSP), pages 1–6. IEEE, 2017.
34. Asif, Sohaib, et al. "AI-Based Approaches for the Diagnosis of Mpox: Challenges and Future Prospects." Archives of Computational Methods in Engineering (2024): 1-33.
35. Shahzad, Inzamam, et al. "Enhancing ASD classification through hybrid attention-based learning of facial features." Signal, Image and Video Processing (2024): 1-14.
36. Asif, Sohaib, et al. "SKINC-NET: an efficient Lightweight Deep Learning Model for Multiclass skin lesion classification in dermoscopic images." Multimedia Tools and Applications (2024): 1-27.
37. Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Ranking sentences for extractive summarization with reinforcement learning. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1747–1759. Association for Computational Linguistics, jun 2018.
38. Asif, Sohaib, et al. "Advancements and Prospects of Machine Learning in Medical Diagnostics: Unveiling the Future of Diagnostic Precision." Archives of Computational Methods in Engineering (2024): 1-31.
39. Ali Nawaz, Maheen Bakhtyar, Junaid Baber, Ihsan Ullah, Waheed Noor, and A. Basit. Extractive text summarization models for urdu language. Inf. Process. Manag., 57:102383, 2020.
40. Makbule Ozsoy, Ilyas Cicekli, and Ferda Alpaslan. Text summarization of Turkish texts using latent semantic analysis. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 869–876, Beijing, China, aug 2010. Coling 2010 Organizing Committee. URL https://www.aclweb.org/anthology/C10-1098.
41. Khan, Umer Sadiq, and Saif Ur Rehman Khan. "Boost diagnostic performance in retinal disease classification utilizing deep ensemble classifiers based on OCT." Multimedia Tools and Applications (2024): 1-21.
42. Dai, Qianwei, et al. "Image classification for sub-surface crack identification in concrete dam based on borehole CCTV images using deep dense hybrid model." Stochastic Environmental Research and Risk Assessment (2024): 1-18.
43. Khan, Saif Ur Rehman, and Sohaib Asif. "Oral cancer detection using feature-level fusion and novel self-attention mechanisms." Biomedical Signal Processing and Control 95 (2024): 106437.
44. Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. CoRR, abs/1705.04304, 2017. URL http: //arxiv.org/abs/1705.04304.
45. Ji Pei, Rim Hantach, Sarra Ben Abbès, and Philippe Calvez. Towards hybrid model for automatic text summarization. In 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 987–993. IEEE, 2020.
46. Raza, A.; Meeran, M.T.; Bilhaj, U. Enhancing Breast Cancer Detection through Thermal Imaging and Customized 2D CNN Classifiers. VFAST Trans. Softw. Eng. 2023, 11, 80–92.
47. Khan, S.U.R.; Asif, S.; Bilal, O.; Ali, S. Deep hybrid model for Mpox disease diagnosis from skin lesion images. Int. J. Imaging Syst.Technol. 2024, 34, e23044.

48. Khan, S.U.R.; Zhao, M.; Asif, S.; Chen, X.; Zhu, Y. GLNET: Global–local CNN's-based informed model for detection of breast cancer categories from histopathological slides. J. Supercomput. 2023, 80, 7316–7348.
49. Khan, S.U.R.; Zhao, M.; Asif, S.; Chen, X. Hybrid-NET: A fusion of DenseNet169 and advanced machine learning classifiers for enhanced brain tumor diagnosis. Int. J. Imaging Syst. Technol. 2024, 34, e22975.
50. Maida Shahid, Summra Saleem, Aniqa Dilawari, and Usman Ghani Khan. A rewriter model for urdu document concision with neural word embeddings. Urdu News Headline, Text Classification by Using Different Machine Learning Algorithms, page 39, 2019.
51. Farooq, M.U.; Beg, M.O. Bigdata analysis of stack overflow for energy consumption of android framework. In Proceedings of the 2019 International Conference on Innovative Computing (ICIC), Lahore, Pakistan, 1–2 November 2019; pp. 1–9.
52. Xin Wan, Chen Li, Ruijia Wang, Ding Xiao, and Chuan Shi. Abstractive document summarization via bidirectional decoder. In International Conference on Advanced Data Mining and Applications, pages 364–377. Springer, 2018
53. Dingding Wang and Tao Li. Weighted consensus multi-document summarization. Information Processing & Management, 48(3):513–523, 2012.
54. Khan, S.U.R.; Raza, A.;Waqas, M.; Zia, M.A.R. Efficient and Accurate Image Classification Via Spatial Pyramid Matching and SURF Sparse Coding. Lahore Garrison Univ. Res. J. Comput. Sci. Inf. Technol. 2023, 7, 10–23.
55. Divakar Yadav and Vimal Kumar K. An improvised extractive approach to hindi text summarization. Volume 339, 09 2015. ISBN 978-81-322-2249-1. doi: 10.1007/978-81-322-2250-7_28.
56. Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. Graph-based neural multi-document summarization. In Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pages 452–462, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/ K17-1045. URL https://www.aclweb.org/anthology/K17-1045.
57. Jen-Yuan Yeh, Hao-Ren Ke, and Wei-Pang Yang. Chinese text summarization using a trainable summarizer and latent semantic analysis. Volume 2555, pages 76–87, 12 2002. ISBN 978-3-540-00261-1. doi: 10.1007/3-540-36227-4_8.
58. Hyeongu Yun, Yongkeun Hwang, and Kyomin Jung. Improving context-aware neural machine translation using self-attentive sentence embedding. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 9498–9506, 2020.
59. Yoko Konno Katsushi Matsubayashi Yuuki Iwasaki, Akihiro Yamashita. Japanese abstractive text summarization using bert.
60. Akhtar, F., Li, J., Yan, P., Imran, A., Shaikh, G. M., & Xu, C. (2020). Exploiting ensemble classification schemes to improve prognosis process for large for gestational age fetus classification. In 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC) (pp. 1455-1459). IEEE.
61. Imran, A., Li, J., Pei, Y., Akhtar, F., Mahmood, T., & Zhang, L. (2021). Fundus image-based cataract classification using a hybrid convolutional and recurrent neural network. The Visual Computer, 37, 2407-2417. Springer Berlin Heidelberg.