



EAST WEST UNIVERSITY

**Domain-Independent, User-centric Text Classification,
and Clustering Framework**

*A capstone project report to be submitted in partial fulfillment of the
requirements for the degree*

of

Bachelor of Science in Computer Science and Engineering

by

Sumona Yeasmin
ID: 2018-2-60-062

Nazia Afrin
ID: 2018-2-60-023

Kashfia Saif
ID: 2018-2-60-001

Under the supervision of

Dr. Mohammad Rezwanul Huq

Associate Professor

Department of Computer Science and Engineering

East West University

Dhaka, Bangladesh

30 June 2022

DECLARATION

Project Title Domain-Independent, User-centric Text Classification, and Clustering Framework

Authors Sumona Yeasmin, Kashfia Saif, and Nazia Afrin

Student IDs ID: 2018-2-60-062, ID: 2018-2-60-001, and ID: 2018-2-60-023

Supervisor Dr. Mohammad Rezwanul Huq

We, hereby, declare that the work presented in this capstone project is the outcome of the investigation performed by us under the supervision of Dr. Mohammad Rezwanul Huq, Associate Professor, Dept. of Computer Science and Engineering, East-West University, Dhaka, Bangladesh. We also declare that no part of this study has been or is being submitted elsewhere for awarding any degree or diploma.

Sumona Yeasmin
ID : 2018-2-60-062
(Signature)

Countersigned

Kashfia Saif
ID : 2018-2-60-001
(Signature)

Dr. Mohammad Rezwanul Huq
Associate Professor
(Supervisor)

Nazia Afrin
ID : 2018-2-60-023
(Signature)

Date : 30 June 2022

Dept. of Computer Science and Engineering
East West University
Dhaka, Bangladesh

LETTER OF ACCEPTANCE

This is to certify that the capstone project entitled **Domain-Independent, User-centric Text Classification, and Clustering Framework**, submitted by **Sumona Yeasmin** (ID: 2018-2-60-062), **Kashfia Saif** (ID: 2018-2-60-001), and **Nazia Afrin** (ID: 2018-2-60-023) are undergraduate students of the **Dept. of Computer Science and Engineering** has been examined. Upon recommendation by the examination committee, we hereby accorded our approval as the presented work and submitted report fulfills the requirements for its acceptance in partial fulfillment for the degree of Bachelor of Science in Computer Science and Engineering.

Dr. Mohammad Rezwanaul Huq
Associate Professor

Dept. of Computer Science and Engineering
East West University
Dhaka, Bangladesh

Dr. Taskeed Jabid
Chairperson, Associate Professor

Dept. of Computer Science and Engineering
East West University
Dhaka, Bangladesh

Date : 30 June 2022

ACKNOWLEDGEMENTS

In the name of Allah, the Most Merciful, and the Most Compassionate, Alhamdulillah, all praises to Allah for the strengths and His blessing in completing this capstone project.

First and foremost, we would like to express our deep and sincere gratitude to our research supervisor, Dr. Mohammad Rezwanaul Huq, for allowing us to conduct research and providing invaluable guidance throughout this work. His dynamism, vision, sincerity, and motivation have deeply inspired us. He has taught us the methodology to carry out the work and to present the works as clearly as possible. It was a great privilege and honor to work and study under his guidance.

We are greatly indebted to our honorable teachers of the Department of Computer Science and Engineering at the East West University who taught us during our study. Without any doubt, their teaching and guidance have completely transformed us to the persons that we are today.

We want to thank S M Keramat Ali (Assistant Engineer at the Department of Transportation in California, USA. MA in Language & Linguistics, University of Dhaka) For supporting us in result validation analysis of the classification on Daily star Dataset.

We are extremely thankful to our parents for their unconditional love, endless prayers and caring, and immense sacrifices for educating and preparing us for our future. We would like to say thanks to our friends and relatives for their kind support and care.

Finally, we would like to thank all the people who have supported us to complete the project work directly or indirectly.

Sincerely -
Sumona Yeasmin
Kashfia Saif, and
Nazia Afrin

Dept. of Computer Science and Engineering
East West University
Dhaka, Bangladesh.

ABSTRACT

Traditional text document clustering and classification methods represent documents with uncontextualized word embeddings and vector space models. Recent text clustering and classification techniques often rely on word embeddings as a transfer learning component. We have explored the existing text document clustering and classification methodologies and evaluated their strengths and weaknesses. We have started with models based on Bag of Words and shifted towards transformer-based architectures. We have concluded that transformer-based embedding will be necessary to capture the contextual meaning. BERT's (Bidirectional Encoder Representations from Transformers) architecture produces robust word embeddings analyzing both from left to right and proper context. Several classification and clustering algorithms have been applied to the word embeddings of the pre-trained state-of-art BERT model. This research has conducted experimental analysis on both classification and clustering algorithms to examine the output on two different datasets. The result analysis of the classification algorithm shows that the random forest classifier obtains around 75% accuracy which is higher than the decision tree and k-nearest neighbor (KNN) algorithms. Furthermore, the obtained results have been compared with existing similar work and show up to 50% improvement in accuracy. The clustering analysis shows that the K-Means has obtained a maximum of 0.654 in Dunn index measurement and 0.135 in Silhouette coefficient, and DBSCAN has obtained a maximum of 0.115. Our capstone project introduces a novel domain-independent, user-centric text clustering, and classification framework. With a Multi-domain text clustering search system, an agent will perform based on user behavior with the user profile. Users will explore document collections by selecting multiple repositories. Users can upload an un-categorized document, and the developed Framework will find similar documents. The developed prototype provides context to the similarity and also finds similar documents within the same domain based on user preferences.

Keywords: Natural Language Processing, Classification, Clustering, Transformer-based embedding, Contextual Similarity, Cluster-Domain Mapping.

Table of Contents

Declaration	1
Letter of Acceptance	2
Acknowledgement	3
Abstract	4
Table of Contents	5 – 6
List of Figures	7
List of Tables	8
1 Introduction	9 – 13
1.1 Research Question	10 -11
1.2 Research Objective	12
1.3 Focus and Contributions	13
1.3.1 Key Contributions	13
1.4 Organization of the Book	14
2 Background & Related Work	15 -27
2.1 Existing Methodologies	16- 24
2.1.1 Sparse Vector Representations (Bag of Word model)	16 – 18
2.1.2 Dense Vector Representations	18 -20
2.1.3 Transformer-based Vector Representations	20 – 24
2.2 Literature Review	25 -26
2.3 Limitations	27
3 Materials and Methods	28 – 48
3.1 Dataset	29 – 30
3.1.1 Reuters Dataset	29 – 30
3.1.2 Daily Star Dataset	30
3.2 Dataset Preprocessing	31 -32
3.3 Design & Implementation	33 -38

3.4	Project Management Features	39 – 45
3.4.1	Business Model Canvas	39– 40
3.4.2	Work Breakdown Structure	40 – 41
3.4.3	Resource Allocation	42 – 43
3.4.4	Critical Path Method	43 -44
3.4.5	Infrastructure Cost	45
3.4.6	Break-Even Point Calculation.....	45
3.5	Test Structure	46 – 47
3.6	Materials and Devices	48
4	Results	49 – 61
4.1	Classification of Reuter Dataset	50 – 52
4.2	Classifier Analysis on Daily Star Dataset	53 – 54
4.3	Clustering on Daily Star Dataset	55 –60
4.4	Key Findings	61
5	Conclusion & Future Work	62
	Reference	63 – 64
	Appendix	65 -76
	Appendix A	65 – 67
	Appendix B	68 – 76

List of Figures

1	Flow Chart of this Book	14
2	Hierarchical Visualization of Existing Methods on Word Embeddings	16
3	Representation of Word in Count Vectorizer	17
4	Representation of Word in One Hot Vectorizer	18
5	Skip Gram Model	19
6	CBOW Model	19
7	Transformer Architecture	21
8	Transformer Encoder Architecture	21
9	Reuter Dataset Size vs Execution Time Analysis	29
10	Pre-Processing Diagram	31
11	Workflow Diagram of the Framework	34
12	Prototype Framework	34
13	Classification and Clustering Framework	36
14	Business Model Canvas	40
15	Work Breakdown Structure(WBS) of Doc Searcher	41
16	Critical Path Method(CPM)	44
17	Reuter Dataset Analysis	51
18	Different Classifier on Reuters Dataset Size vs Accuracy	51
19	Performance Comparison of Capstone's Accuracy with the Targeted Research Work on Reuters Dataset	52
20	Daily Star Dataset and Classifier Analysis	53
21	Mapping Clusters with Domain of KMeans Result	57
22	Mapping Clusters with Domain of DBSCAN Result	58

List of Tables

1 Reuter Dataset Description(for result validation)	29
2 Reuter Dataset Description(for performance evaluation)	30
3 Daily Star Dataset Description	30
4 Resource Allocation	42 – 43
5 Employee’s Salary Estimation	43
6 Infrastructure Cost	45
7 Test Structure	46– 47
8 Classifier Algorithm performance analysis on Reuter Dataset	50
9 Performance Comparison with the Target Research	52
10 Classifier Algorithm performance analysis on Daily Star Dataset	53
11 Test Data analysis of Daily Star Dataset	54
12 Unseen Test Data class label comparison between classifier predictions and System generated Vs. classifier predictions and Domain experts identified	54
13 KMeans and KMedoid Cluster Quality Analysis	55
14 DBSCAN Cluster Analysis	56
15 Confusion Matrix of KMeans on Daily Star Dataset	58
16 Confusion Matrix of DBSCAN on Daily Star Dataset	59
17 Comparson between KMeans and DBSCAN Clustering Result	59

1

Introduction

Natural language processing (NLP) is the ability of a computer program to understand human language. NLP consists of two significant steps, which are pre-processing and implementing algorithms. Pre-processing consists of two techniques which are syntax (arrangements of words in a sentence to make grammatical sense) and semantic (the use of and meaning behind words) analysis. Text document clustering has become one of the most highly essential techniques. It has a wide range of applications such as document organization, content identification, information retrieval, and similar document searching [1]. Typically, descriptors or context are extracted from the document first. Then based on that same context, records are grouped and kept together, making it more convenient and easier to recommend more documents underlying the same group. Documents cannot be categorized or clustered using a traditional clustering or classification algorithm which requires the input to be a fixed-length feature vector. That's why it is crucial to figure out how we represent or vectorize large text documents. Algorithms only can work with numeric values. A challenge arises to represent a vast amount of text in numerical form, capturing a particular text's semantic meaning and syntax. We introduce a multi-domain text clustering and classification framework by embedding the state-of-art BERT (Bidirectional Encoder Representations from Transformers) model, which has mitigated these drawbacks. We have used the powerful multi-head mechanism of BERT to vectorize text documents with contextual sentence embeddings. After the vectorizations, we have clustered and classified the documents with algorithms such as K-Means and DBSCAN for clustering and Decision Tree, Random Forest, and K-nearest neighbors for classification.

1.1 Research Question

The core of each systematic inquiry is a set of research questions. The core focus of the systematic study is research questions. Asking the right questions when conducting research can help you collect relevant and insightful information that positively influences your work.

Research questions help identify the minor question needed to answer and keep the focus of a study. An appropriate set of research questions can guide the study's goal.

In this section, we have demonstrated the solution approach to our document classification clustering problem and the development of such a framework that meets specified needs for public health and safety, cultural, societal, and environmental considerations. First, we mention the research questions.

How to design and develop a domain-independent, user-centric text clustering framework and assess the impact under cultural, societal, environmental, ethical, and legal frameworks?

RQ1. How to apply and integrate new and previously acquired mathematics, science, and engineering knowledge to address the challenges associated with the Capstone Project? (PO1)

RQ2. What relevant domains needed to be explored, and how to define the problems and formulate the objectives of the Capstone Project? (PO4)

RQ3. How to analyze the various aspects of the objectives of the Capstone Project to design an efficient solution? (PO2)

RQ4. How to design and develop solutions for the capstone project that meet public health and safety, cultural, societal, and environmental considerations? (PO3)

RQ5. Which modern engineering and IT tools are required, and how to apply them for designing and developing the solution for the Capstone Project? (PO5)

RQ6. How to assess and address societal, health, safety, legal, and cultural aspects related to implementing the capstone project? (PO6)

RQ7. How to assess and address the sustainability impact of the capstone project in societal and environmental contexts? (PO7)

RQ8. Which professional and engineering ethical principles and practices should be followed for the implementation of the capstone project? (PO8)

RQ9. Which practices should be followed to perform effectively as an individual and a team member to accomplish the objectives of the Capstone Project? (PO9)

RQ10. Which practices should be followed to make effective deliverables? (PO10)

RQ11. How to apply software engineering principles and practices to the development life cycle of the Capstone Project and conduct economic analysis and cost estimation in case of a real-life deployment of the solution of the Capstone Project? (PO11)

RQ12. Which independent and life-long learning are acquired throughout the entire process of the design and development of the Capstone Project? (PO12)

1.2 Research Objectives

The outcomes wanted to attain by doing research are known as research objectives. Developing good research objectives can assist in meeting its overall goals. Research objectives aim to guide the whole research process, from data collection to analysis to conclusions. Study objectives also aid in focusing the research, identifying relevant factors, showing the research process, and pinpointing the major focus of the research.

In this section, the objectives of the research have been defined. The most significant objective of the Capstone Project is to provide a novel and unique text clustering framework with the following characteristics.

Domain-Independent: One of the objectives of this Capstone Project is to deliver a domain-independent text clustering framework that can be effectively used for different domains such as politics, sports, health, science, etc.

User-centric approach: Another objective of the Capstone Project is to allow the text clustering framework to learn the user preferences by creating a user profile to generate better clustering depending on the users' past interactions with the system.

Contextual Analysis: A particular political domain may have different dimensions or contexts. One of the objectives of our Capstone Project is to understand the context within an environment and deliver the results based on that context. Suppose a sports journalist wants to retrieve the content on uprising cricketers. We expect that our novel Framework can understand the context of the document and then provide better results to end-users.

Contextual Similarity: Many robust algorithms can find structural similarities. However, we focus on formulating a novel method to find the contextual similarity of documents by considering the existing methodologies and improving them incrementally. It can be regarded as another objective of the Capstone Project.

Recommender Systems: Since we propose a complete text classification and clustering framework, we have included a text recommender system based on the context and user profile. It could help our end-users to find more meaningful results.

Cost-effective Framework: We plan to follow the Agile software process model to make the complete Framework more cost-effective and accommodating to the changing requirements.

Environment-friendly: The solution of the Capstone Project should be environment-friendly. Our proposed Framework maintains a large corpus of documents and electronically provides results. The proposed Framework is a cloud-based service. The model is fully paperless. Therefore, it helps reduce manual work through pen and paper, reducing carbon emissions modestly.

1.3 Focus and Contributions

A web-based application where users can upload documents from any domain and later, they can search for relevant documents. This capstone project has introduced a novel extensive text clustering method using BERT's contextual word embeddings.

This study focuses on using state-of-the-art tools and technologies and bringing them under the same umbrella. We have focused on using transformer-based architecture BERT, which is mostly used for next word prediction, but we have applied it in a different approach which is text classification and clustering. The study of embedding schemes of different models concludes that the transformer-based architectures are better suited for generating contextual word embeddings.

For classification, experiments show that the random forest classifier obtains the highest accuracy than the decision tree and k-nearest neighbor (KNN) algorithms. Furthermore, the obtained results have been compared with existing work and show up to 50% improvement in accuracy [2].

Several clustering algorithms have been applied over the embedding of BERT to cluster large text documents based on contextual meaning. Experiments show that the K-Means has obtained a maximum of 0.654 in Dunn index measurement and 0.135 in Silhouette coefficient, and DBSCAN has obtained a maximum of 0.115 in Silhouette coefficient.

1.3.1 Key Contributions:

The framework focuses on the following:

- Making it more convenient and easier to retrieve documents.
- Contextual word embeddings are better derived with transformer-based architectures.
- The powerful multi-head mechanism of BERT is used to vectorize text documents with contextual sentence embeddings.
- Applied several clustering algorithms over the pre-trained BERT-generated vectors to enhance our understanding of text document clustering.
- Using all the state-of-the-art tools and technologies and bringing them under the same platform.

1.4 Organization of the Book

The structure of our book is as follows. We are starting with Chapter 1, which introduces the project's principal goal. In Chapter 2, we'll summarize some existing work that's relevant to our research topic and describes the current work on the relevant subject. It also explains the study's major focus and limitations, as well as the research aims. Chapter 3 introduces the materials used in this project as well as research methodologies such as design and implementation. Project management features are also included in Chapter 3. In Chapter 4, the model's results will be analyzed and evaluated. In Chapter 5, the study concludes with a consideration of future work. In figure 1, we have shown the flow chart of this book.

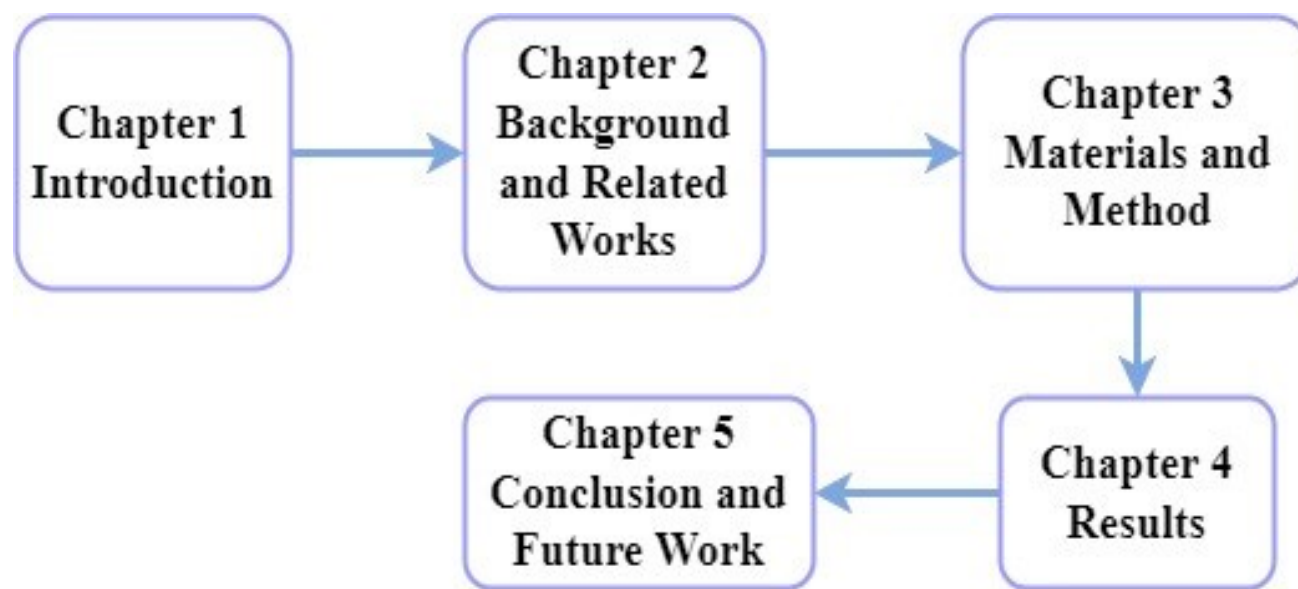


Figure 1: Flow Chart of this Book

Background and Related Work

Document classification and Clustering have been central research areas in the field of Natural language processing. Classification and Clustering of the large text document are pretty complex as the contextual embedding of words is needed to preserve the document's meaning. The background study of this research started with the extensive methods of word vectorization and moved toward Transformer based embeddings. Several background studies have shown that the Transformer based architectures possess robust word embedding schemes, which are ideal for preserving the contextual meaning of a text. Literature review establishes the current state of knowledge in the field of the issue or topic a study is working about. After exploring current state-of-art word embeddings methodologies, this research has studied the existing related work to identify the strength and weaknesses of the reviewed literature. A related work-study can survey the literature in the chosen area of the research. It synthesizes the information in that literature into a summary.

2.1 Existing Methodologies

Figure 2 shows the hierarchical visualization of existing methods on word embeddings. To our knowledge of existing methodologies of word embeddings, we have observed three types of embeddings, which we have explored in the following terms:

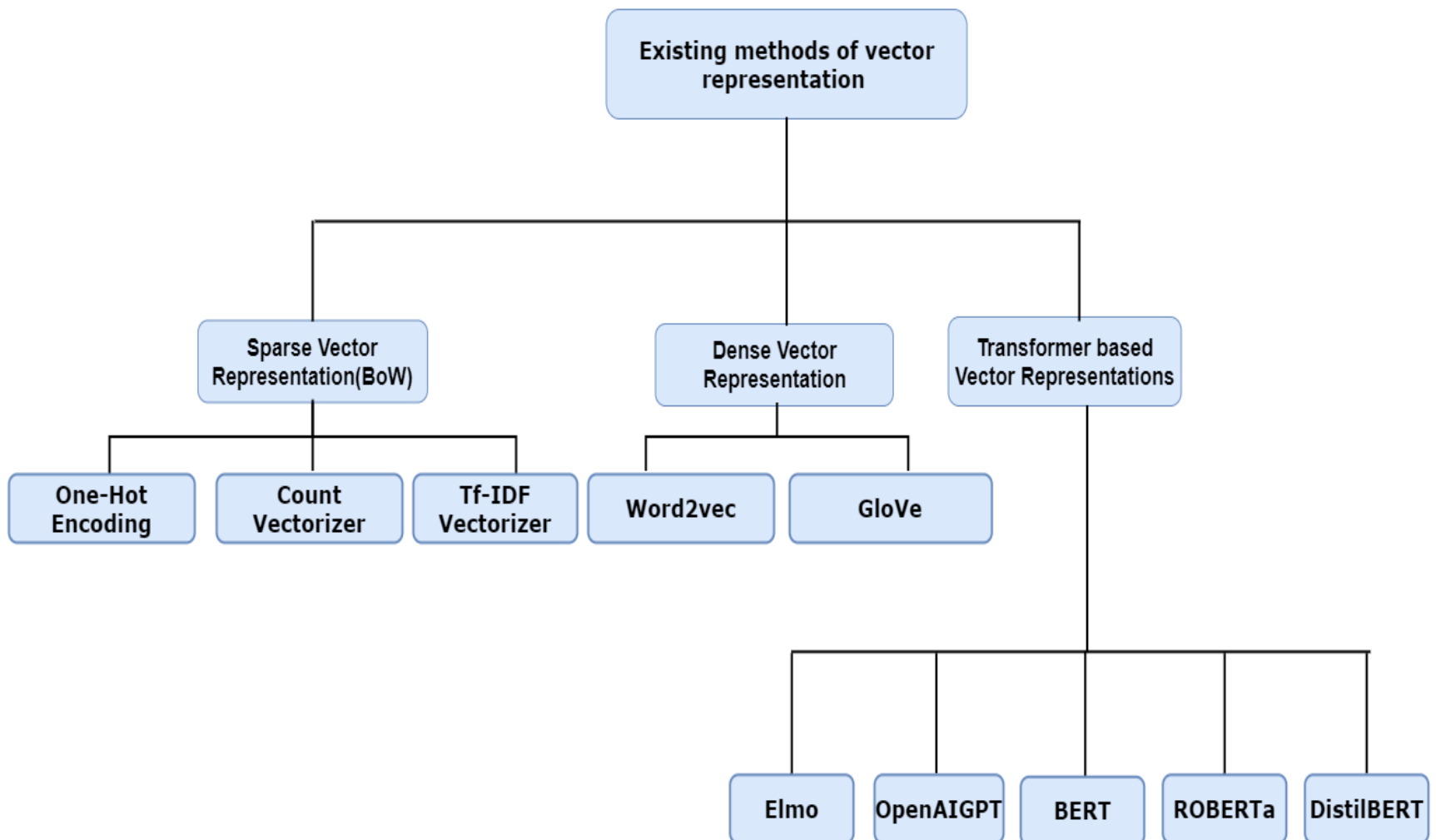


Figure 2: Hierarchical visualization of existing methods on word embeddings

2.1.1 Sparse Vector Representations (Bag of Word model)

Count Vectorizer

Count Vectorizer takes the count value of each occurring word and transforms them into vectors. It transforms a given text into a vector based on each word's frequency (count) in the entire text. Count Vectorizer is used to convert a collection of text documents to a vector of term/token counts. Example Input Document :

doc=[" One Cent, Two Cents, Old Cent, New Cent: All About Money"]

Representation in Count Vectorizer:

	about	all	cent	cents	money	new	old	one	two	
doc	1	1	3	1	1	1	1	1	1	In theory

↓

index	0	1	2	3	4	5	6	7	8	
doc	1	1	3	1	1	1	1	1	1	In practice

Figure 3: Representation of words in count vectorizer

Tf-IDF Vectorizer

TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. TF-IDF (term frequency-inverse document frequency) is a vector representation scheme. TF-IDF calculates the importance or the relevance of a word document within a collection of documents. The calculation is done by multiplying two metrics:

- frequency of a word in the document
- Inverse document frequency of that particular word across a set of documents.

The higher the value of TF-IDF, the more relevant that word is in that particular document.

The calculation of TF-IDF:

First, we calculate the term frequency of each word given the formula:

$$idf(t) = \text{occurrence of } t \text{ in } d \text{ document}$$

The IDF value is calculated by dividing the total number of documents with the term frequency value:

$$idf(t) = \frac{N}{tf(t,d)} \quad (N = \text{number of documents})$$

In the case of a large corpus, the IDF value also becomes vast. To avoid the effect, we take the log of IDF :

$$idf(t) = \log \left(\frac{N}{(tf(t,d)+1)} \right)$$

Finally, we get the value of TF-IDF by multiplying the term frequency and Inverse Term frequency together:

$$tf-idf(t,d) = tf(t,d) * \log \left(\frac{N}{(tf(t,d)+1)} \right)$$

One Hot encoding

One hot encoding is another method of converting data for an algorithm and better prediction. We convert each categorical value into a new flat column with one-hot and assign a binary value of 1 or 0 to those columns [3]. Marks a particular vector index with a value of true (1) if the token exists in the document and false (0) if it does not (Figure: 3). In other words, each element of a one-hot encoded vector reflects either the presence or absence of the token in the described text [4].

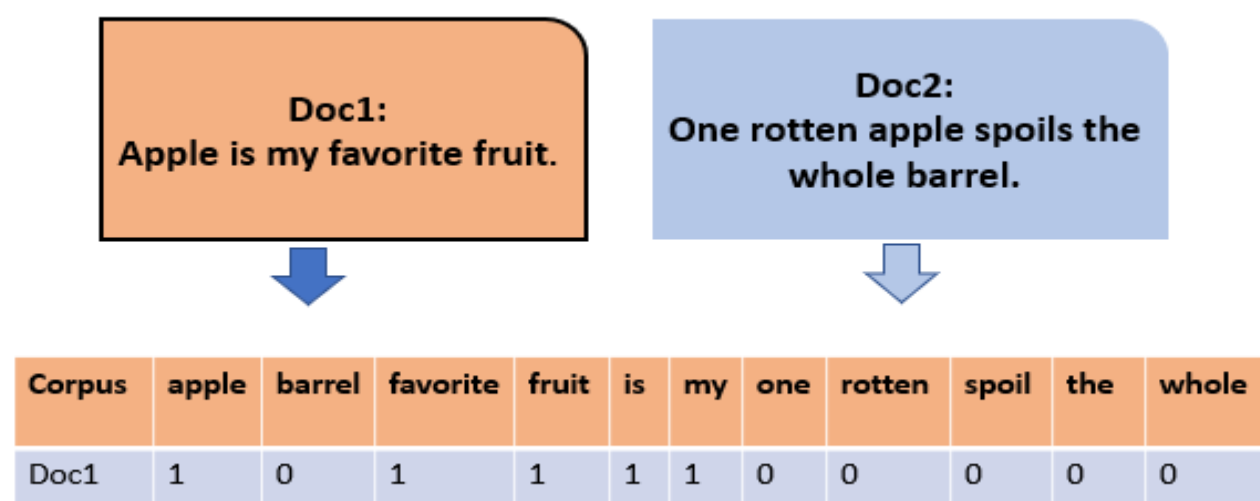


Figure 4: Representation of words in one hot vectorizer

The limitations of Sparse Vector representations

Despite being very popular vectorization techniques, sparse vector representations possess several limitations. Both count vectorizer and Tf-IDF cannot identify the relationships between words in terms of linguistic similarity and word importance for analysis, which signifies that sparse vector representations cannot capture the semantic meaning and the sequence of words in a sentence.

2.1.2 Dense Vector Representations

Word2vec

Word2vec is a statistical method for efficiently learning a standalone word embedding from a text corpus [5]. The word2vec objective function causes the words in similar contexts to have similar embeddings. The word2vec is a six technique in NLP that uses a neural network model to learn word associations from a large text corpus. Once trained, such a model can detect synonymous words or suggest more words for a partial sentence. Word2vec uses a window size that determines the number of context or input words. As the name implies, word2vec represents each word with a particular list of numbers called a vector [6]. Word2vec can obtain using two methods:

2.1 Existing Methodologies

- Skip-gram: Skip-gram feed-forward neural network takes the target word as input and compares the context words as output node values.
- CBOw(continuous bag of words): CBOw feed-forward neural network iterates context word vectors as input node values and predicts the target word vector.

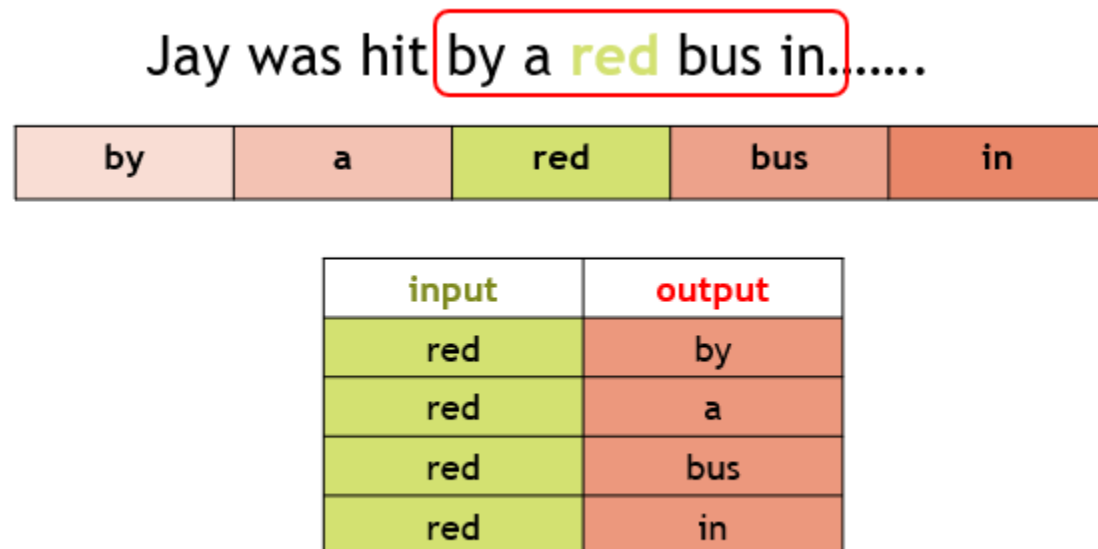


Figure 5: Skip gram model

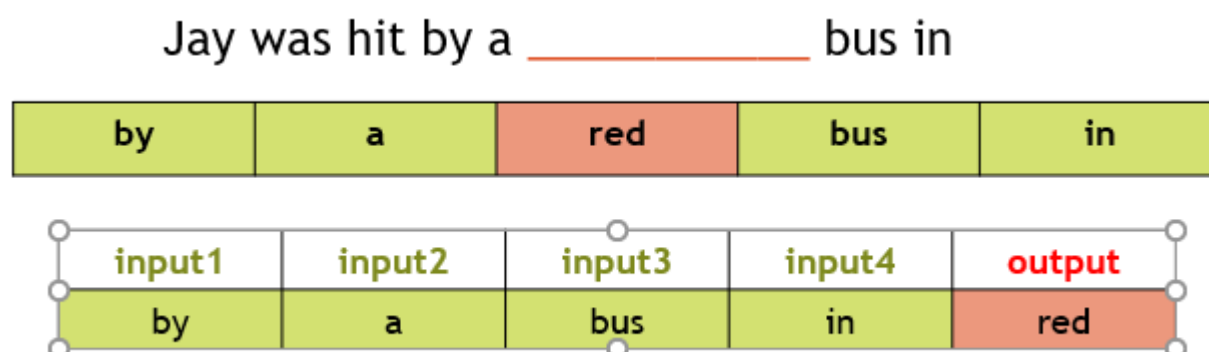


Figure 6: CBOw model

GloVe

The GloVe is an unsupervised algorithm for vector representations for words. The training performs on aggregated global word-word co-occurrence statistics from a corpus, and the resulting drawings show interesting linear relations of the word vector space. The gloVe is the more delicate structure of the word vector space. It examines not the scalar distance between word vectors but their various dimensions of difference. GloVe and Word2Vec vectorize text so that each word in a text gets its multidimensional vector, and each size captures the different characteristics of that particular word.

Dense vector representations shift from sparse vector models as they can somewhat catch the similarity of words, meaning they can represent words such that semantically similar words end up close to each other in the embedding space. But they cannot capture the contextual meaning of the whole paragraph. The main challenge of sparse vector representations is their inability to handle unknown or out-of-vocabulary words. If the models have not seen a word

2.1 Existing Methodologies

Previously, they do not know how to build a vector for it, resulting in a random vector, which is far from ideal. For example, Word2vec represents every word as an independent vector, even though many are morphologically similar. So, they are not applicable for initializing state-of-the-art architectures [7]. Another limitation of sparse vector representations is that they are not bidirectional. Models can take account of a sequential text from either left to right or right to left, averts capturing the complete contextual meaning of a document or a paragraph.

2.1.3 Transformer-based Vector Representations

The sparse and dense vector representations have many limits. We explore the current time's state of art transformer-based architecture models to address these limitations. Transformers have robust word embedding schemes that capture a word's semantic and contextual meanings. Transformers' idea is to completely handle the dependencies between input and output with attention and recurrence. Attention mechanisms have become an integral part of captivating sequence modeling and transduction models in various tasks, allowing the modeling of dependencies without regard to their distance in the input or output sequences [8] [9] . The Transformer architecture consists of a multi-head attention mechanism and feed-forward neural network (Figure: 6). The Transformers follows this overall architecture using stacked self-attention and pointwise, fully connected layers for the encoder and decoder.

Transformer Architecture

The transformer architecture shown in Figure 7 has an encoder-decoder structure (figure 8).

The output of the encoder is a continuous vector representation of inputs.

Encoder: The encoder comprises a stack of $N = 6$ identical layers. Each layer has two sublayers. The first is a multi-head self-attention mechanism; The second is a simple, position-wise, fully connected feed-forward network.

Decoder: The decoder comprises a stack of $N = 6$ identical layers.

In addition to the two sub-layers in each encoder layer, the decoder inserts a third sub-layer, which performs multi-head attention over the output of the encoder stack [10]. BERT base has 12 layers, 12 attention heads, and 110 million parameters. BERT Large has 24 layers, 16 attention heads, and 340 million parameters. Transformer architectures apply a three-step Text Processing:

- Position embedding: position of a word in the sentence.
- Segment embedding: can take penalty pairs as inputs. It learns a unique embedding for first and second sentences.
- Token embedding: The embeddings learned for the specific token from the WordPiece token vocabulary.

2.1 Existing Methodologies

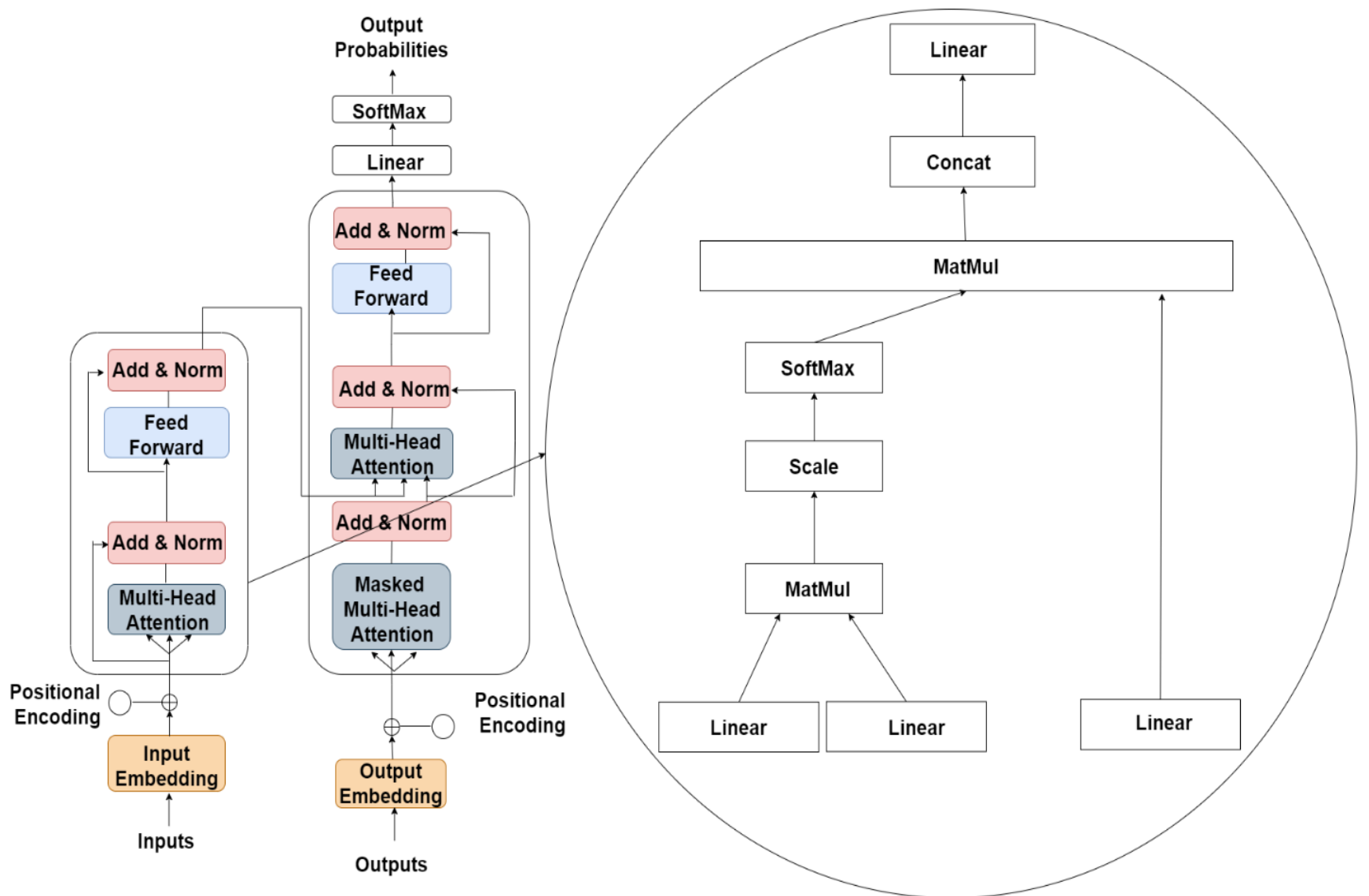


Figure 7: Transformer Architecture

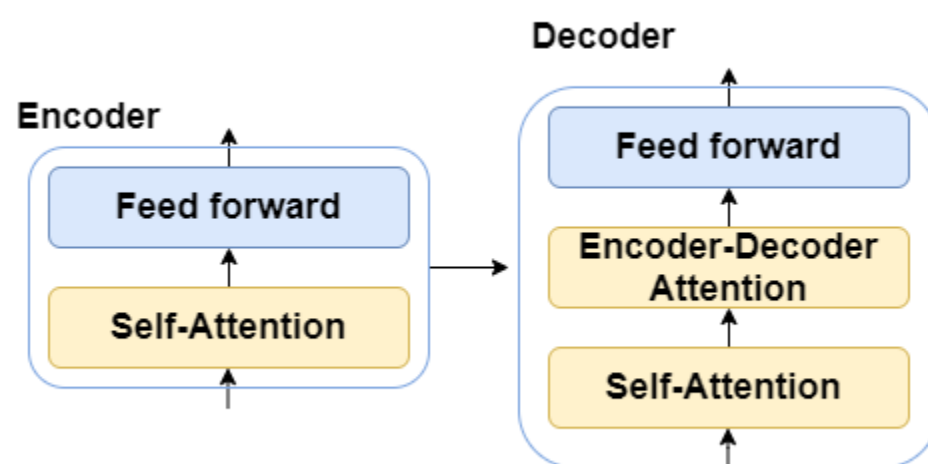


Figure 8: Transformer Encoder-Decoder Architecture

Transformers have a self-attention mechanism, which can map a query and a set of key-value pairs to an output, where the query, keys, values, and production are all vectors [10]. The steps followed in single self-attention are:

Producing the Key, Query, and value Vector

Self-attention calculation includes creating three vectors (Query, Key, Value) from each of the encoder's input vectors. So, for each word, we create a Query vector, a Key vector, and a Value vector. In encoder-decoder attention layers, the queries come from the previous decoder layer, and the memory keys and values come from the output of the encoder [10]. These Query, Key, and Value vectors are created by passing the input embedding merged with the positional encoding vector of each word to a linear layer.

Multiplication of query and key vector

This step includes calculating a score. We need to find the score for each word against the targeted word. The score determines how much focus on placing on other parts of the input sentence as we encode a comment at a particular position. The score is calculated by taking the dot product of the query vector with the essential vector of the respective word we're scoring [11]. Next, we divide the scores by the square root of the dimension of the key vectors. This calculation gives more stable gradients. Then the resultant vector is passed through a SoftMax operation. SoftMax normalizes the scores. This SoftMax score determines the importance of a word in a particular position.

Matmul with Value vector

The fifth step consists of multiplying each value vector by the SoftMax score.

Concatenation

Next, we sum up the weighted value vectors. Weighted value vectors produce the output of the self-attention at its position for each word. The concatenated vector then passed through a final linear layer which is now the output of a single attention head called filtered value matrix. BERT base has a total of 8 attention heads that focus on the different linguistic features of a single sentence—like this creating a multi-head attention mechanism. Multi-head attention allows the model to jointly attend to information from different representation subspaces [10]. The encoder stack results from this multi-head attention and passes it through an add and norm layer and then into a feed-forward neural network. The neural network's output is then passed through an add and norm layer, which produces the final result of the encoder. Each position in the encoder can attend to all posts in the previous layer of the encoder [10]. The development of the encoder stack is fed as input into the Decoder stack. Decoder Multi-head takes three inputs: The Key, Value vector from an encoder, the masked multi-head output. Decoder consists of masked multi-head attention, multi-head attention, add and norm layer similar to the encoder stack, and a feed-forward neural network [10]. The output of the decoder is passed to a final linear layer with SoftMax to produce the result.

Elmo

Elmo is a bi-LSTM language model. Elmo stands for Embeddings from Language Model, as the name suggests the deeply contextualized word embeddings created from the Language Models (LM). The feature-based approach, Elmo (Peters et al., 2018a), uses shallowly bi-directional task-specific architectures that include the pre-trained representations as additional features [12]. Elmo uses a bidirectional language model (biLM). Elmo is pre-trained on a large text corpus to learn words (e.g., syntax and semantics) and linguistic context (i.e., to model polysemy). BiLM captures context-dependent aspects of word meaning. Elmo uses the concatenation of forwarding and backward LSTMs [13].

OpenAIGPT

The OpenAI GPT is the successor of the GPT model. OpenAI GPT is a large transformer-based language model, with generative pre-training of a language model on a diverse corpus of unlabeled text, followed by discriminative fine-tuning on each specific task [13]. The fine-tuning approach is Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), which introduces minimal task-specific parameters, all pre-trained parameters. In OpenAI GPT, the authors use a left-to-right architecture, where every token can only attend to previous permits in the self-attention layers of the Transformer. GPT uses a multi-layer transformers decoder [13].

BERT

BERT stands for Bidirectional Encoder Representations from Transformers. BERT is designed to pre-train deep bidirectional representations from an unlabeled text, by jointly conditioning the left, right and proper context in all layers [12]. BERT has two major steps to complete its task Pre-training and fine-tuning. The model is trained on unlabeled data over different pre-training tasks during pre-training. BERT model first comes with pre-trained parameters, and all the parameters are fine-tuned using labeled data from the downstream tasks. Each downstream task has different fine-tuned models, even though initialized with pre-trained parameters.

ROBERTa

A robustly Optimized BERT Pretraining Approach. RoBERTa replication study of BERT pre-training (Devlin et al., 2019) includes a careful evaluation of the effects of hyperparameter tuning and training set size [14]. RoBERTa builds on BERT's language masking strategy and modifies key hyperparameters in BERT, including removing BERT's next-sentence pre-training objective and training with much larger mini-batches and learning rates. Roberta was also trained on the order of magnitude more data than BERT for a more extended time. This allows Roberta representations to generalize better to downstream tasks than BERT [15].

DistilBERT

DistilBERT is a small, fast, cheap, and light transformer model based on the BERT architecture, which is 40% smaller, 60% faster than retains 97% of the language understanding capabilities. Knowledge distillation is performed during the pre-training phase to reduce the size of a BERT model by 40% [16]. DistilBERT is a method to pre-train a smaller general Purpose language representation model, which can be fine-tuned with good performances on a wide range of tasks like its larger counterparts [16]. Focuses on reducing the number of hidden layers of BERT architecture.

2.2 Literature Review

This section has studied the related work to the developed solution and explored their strengths and weaknesses.

Yutong Li et al. [17] have developed a clustering module of large text documents based on BERT's fine-tuning. The authors have obtained the BERT's powerful word embedding capable of preserving the semantic meaning of the Documents. The authors have fine-tuned the BERT model with a weighted scheme and tested their algorithm on the Reuter Dataset.

Wenhao Hu et al. [18] have developed another clustering method of large documents based on an improved K-means algorithm with a Density peak. The authors have used the embeddings of the BERT model and then applied the density peak cluster algorithm to obtain the cluster center. The authors have introduced the BERT model to calculate text similarity, and the advantages of the sentence vector generated by BERT are used in the paper.

Haoxiang Sh et al. [19] have proposed Document Clustering Based on BERT with Data Augmentation. The authors stated self-supervised contrastive learning (SCL) as well as few-shot contrastive learning (FCL) with unsupervised data augmentation (UDA) for text clustering. The authors have turned the BERT model based on their learning methods. Furthermore, the authors of this paper have used learned latent representations to perform clustering methods. Authors have introduced data augmentation methods, back translation, and random masking. The authors' methodologies also show that they can achieve results closer to a supervised machine learning algorithm.

Wei-Cheng Chang et al. [20] have introduced X-BERT, an extreme Multi-label Text Classification using the BERT model. The authors have presented the Extreme multi-label text classification problem and developed a solution to fine-tune BERT. According to the authors of this paper, the proposed model, X-BERT, leverages both the label and input text to build label representations, which induces semantic label clusters to better model label dependencies [20]. At the heart of X-BERT is a procedure to fine-tune BERT models to capture the contextual relations between input text and the induced label clusters. Finally, an ensemble of the different BERT models trained on heterogeneous label clusters leads to our best final model, which leads to a state-of-the-art XMC method [20].

Ashutosh Adhikari et al. [21] have developed DocBERT, a fine-tuned BERT model for Document classification. The authors have used the distilled BERT model small bidirectional

2.2 Literature Review

LSTMs. The authors have improved the baselines for document classification by fine-tuning BERT. Also, the authors have used the knowledge learned by BERT models to enhance the effectiveness of a single-layered lightweight BiLSTM model.

Daniela Godoy et al. [22] Proposed an intelligent agent called a personal searcher. Personal searchers collect documents from the world wide web and filter out the results based on the high probability relevant to the target user. According to the authors, the proposed model gradually builds up a user profile to perform better according to the user's taste. The agent uses a textual case-based reasoning approach to detect specific subjects the user is interested in and organizes them in a hierarchy that defines the user profile [22].

A survey of BERT's performance analysis in [23] has applied four clustering algorithms like k-means clustering, eigenspace-based 17 fuzzy c-means deeply embedded clustering, and improved deeply embedded clustering. According to this study, the embedding of the BERT model yields accurate clustering findings.

2.3 Limitations

The analysis of the studies mentioned above possesses several limitations.

On a technical level, the study of Yutong Li et al. [17] has not included any sentence-level feature in the weighted scheme to improve its overall performance.

Haoxiang Sh et al. [19] have introduced SCL(Self-supervised Constructive Learning). In the proposed method of this study, SCL(Self-supervised Constructive Learning), two texts in the same class may generate two positive pairs for a mini-batch. According to the authors, they have treated these two pairs as negative samples for each other, but this negative sampling limits the performance of their algorithm SCL (Self-supervised Constructive Learning).

In the study by Ashutosh Adhikari et al. [21], the authors have not provided distillation effects over a range of neural network architectures. Also, the authors have not explored the model compression techniques in the context of transformer models.

Overall, we have observed the following limitations of the existing literature.

In Yutong Li et al. [17], Wenhao Hu et al. [18], Haoxiang Sh et al. [19], Wei-Cheng Chang et al. [20] Ashutosh Adhikari et al. [21], the authors have used the BERT's embeddings for Document classification and clustering. There are three limitations we have concluded. Firstly, the authors have not provided any solution to the domain-independent text clustering and classification mechanism.

Secondly, the studies above do not propose a solution to a user-centric approach. Clustering or classification conducted in the studies mentioned above focuses on the performance of the algorithms only. Lastly, No studies have been done for a platform to perform clustering or classification on a cloud-based system such as the developed Framework in these papers.

However, in Daniela Godoy et al. [22], the authors have proposed a user-centric document retrieval system that acts according to the user's needs following user browsing history, but this study has several limitations. The proposed methods in the paper do not use a state-of-art embedding system such as BERT. No modules were developed as a solution for document classification or clustering.

Materials and Method

The underlying plan and reasoning of the research study are referred to as methodology. It entails researching the methods employed in the study and the ideas or concepts behind them to design a strategy tailored to the goals.

This chapter describes the datasets taken for the research. We have taken the dataset we collected from the Daily Star news portal for experimental analysis. The dataset has been created via web scrapping using pythons' beautiful soup library.

Next, this chapter elaborates on the methodologies of the clustering and classification framework and the modules. Further, this chapter also narrates the project management features of the capstone project. Project management shows applying specialized knowledge, skills, tools, and processes to provide a business point of view of the developed prototype. Project management helps the development of software to assist in the improvement of business processes.

3.1 Dataset

3.1 Dataset

3.1.1 Reuter Dataset

The Reuter dataset from Reuters-21578, Distribution 1.0 in Natural Language Toolkit (NLTK) [24], was taken to test our methods for our capstone project. From Reuter's 21 datasets, we have taken those content belonging to only one topic.

From Reuters Dataset, as described in table 1, we have conducted the experiments on four datasets from Reuter. The dataset DS4_v2, DS5_v2, DS8_v2, DS15_v2 consist of 200, 500, 1000, 5000 documents respectively. We have selected this dataset size and topic number similar to Yutong Li et al. [17] to compare the result analysis and validate later.

Table 1: Reuter Dataset Description (for result validation)

Dataset name	Dataset Size	Topic number	Time(sec)
DS4_v2	200	4	372.45
DS5_V2	500	5	784.58
DS8_v2	1000	8	4391.68
DS15_v2	5000	15	8786.54

BERT Execution Time on Reuter Dataset: Figure 9 shows the execution time analysis of the BERT model's vector embedding generation against the Dataset size of the Reuters.

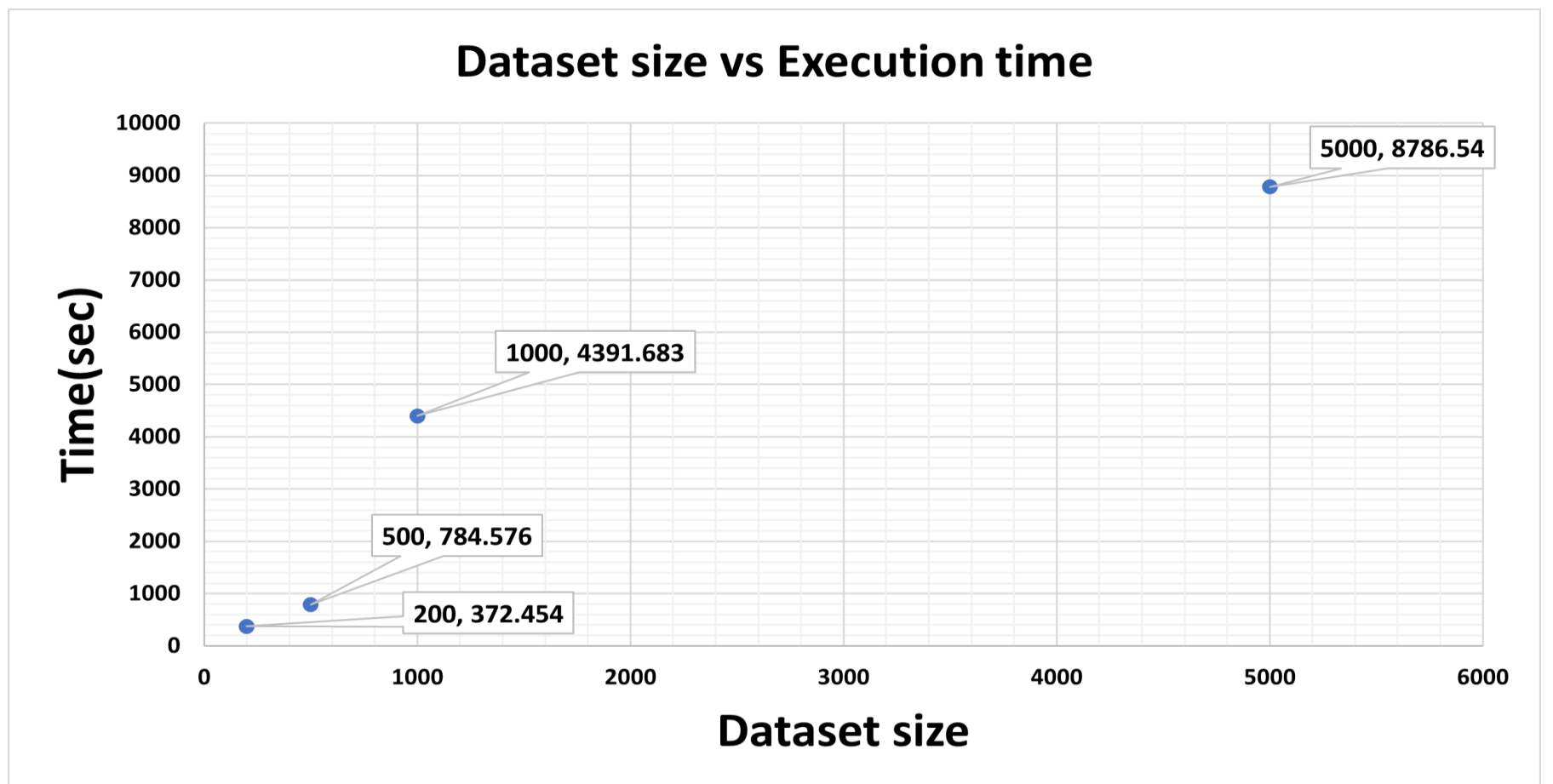


Figure 9: Reuter Dataset Size VS Execution Time Analysis

The capstone project has worked with four Reuters datasets (DS4_v2, DS5_V2, DS8_v2, DS15_v2). We have observed that at first, with the increasing dataset, the time taken by the BERT model increases

3.1 Dataset

at a higher rate. When dataset size rises by 500 to 1000 samples, we see a sharp increase in the time the BERT model takes. Interestingly, the sharpness decreases for a five times higher dataset size, meaning BERT takes less time with the larger dataset. This analysis may signify that the time increment is logarithmic against the size of the dataset.

Table 2 presents the Reuter Dataset Description (for performance evaluation). From the Reuters dataset, we have taken only those contents belonging to topics if their frequency is more than n . We have taken a threshold value of $n=250$ to filter the dataset. This threshold value is assumed to be an experimental value suitable for a balanced dataset. The Most frequent five topics were chosen. Further, the dataset has been sectioned into four categories differing four sizes as below.

Table 2: Reuter Dataset Description (for performance evaluation)

Dataset name	Dataset Size	Topic number
DS1	200	5
DS2	500	5
DS3	1000	5
DS4	5000	5

3.1.2 Daily Star Dataset

In this capstone project, algorithms have been researched using Daily Star dataset. The dataset has been made by web scrapping from Bangladesh's leading online news article. The Beautiful Soup package in Python was used to collect the news material. The documents were tagged with their respective domains to create a complete dataset for experimental purposes. The documents have seven domains: business, entertainment, sports, live living, Tech news, Youth, and Environment. From the scrapped documents, 415 records were taken for experimental analysis.

Table 3: Daily Star Dataset Description

Dataset name	Dataset Size
DS415	415 Documents
DS212	212 Documents

As mentioned in table 3, we also tested our algorithm on the Daily Star dataset. We have collected the dataset via web scraping. For result analysis in this project, we have divided the dataset into two sections DS415, DS212 containing 415, and 212 documents. We have selected Business, Entertainment, Sports, Live living, Tech news, Youth, and Environment a total of seven Domains from the "Daily Star" news portal. The dataset can be found at [25].

3.2 Dataset Pre-processing

Large text files can contain garbage values and characters that should be eliminated before producing word-level embeddings. Therefore, preparing and preprocessing the uncategorized raw text before providing them to the BERT model is a key task for a robust classification result. The following data preprocessing steps (figure 10) were performed over the datasets used for the experiments.

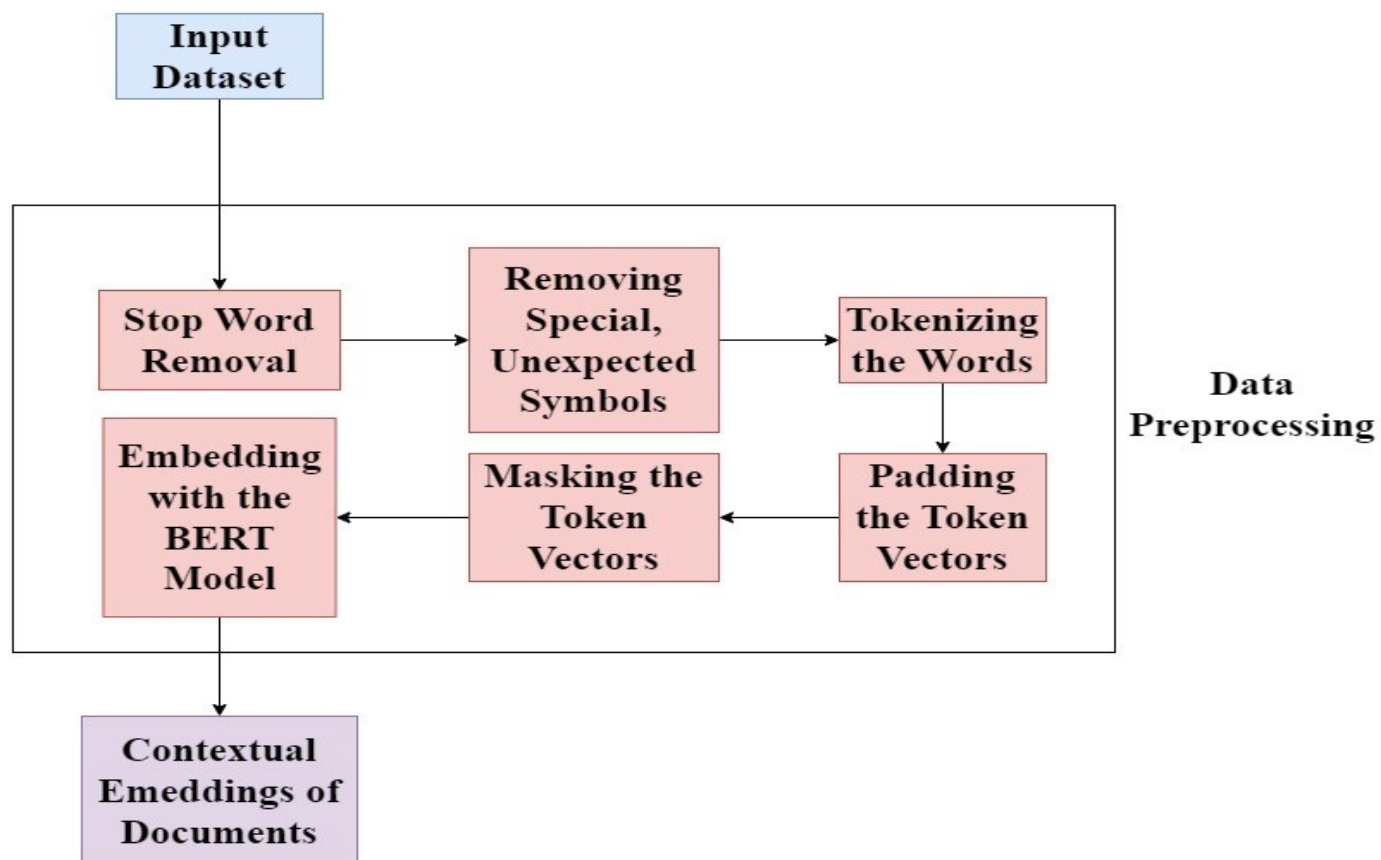


Figure 10: Pre-Processing Diagram

At first, the unwanted symbols and stop words were removed using the python Regex and NLTK libraries. Next, we have tokenized the words as a crucial step for feeding tokens to the BERT model. We divide extensive sentences into words and tokenize them using BERT's tokenizer library. Afterward, padding is done. The padding ensures that all sequences in a batch are the same length. The padding strategies of the BERT model are employed to pad the corpus. For the next step, we have performed masking of the tokens; masking informs sequence processing layers that specific inputs are missing, which helps in better model training. The BERT model was used to apply masking. Lastly, the BERT model was used to produce the embeddings. The BERT model's objective is to get the contextual word embedding that also considers the word positioning and relationship among the words.

1. Stop words removal: Stop words are the most frequent word and they can reduce the integrity of the context of a document. Therefore, Python's NLTK library functions are used to discard the stop word for strong contextual corpus generation. The entire document is divided into words

3.2 Dataset Pre-processing

2. Removing the unwanted symbols: Uncategorized data can come with all kinds of symbols that are not expected during model training. Python's Regex function is used to remove special symbols and characters from the dataset.

3. Tokenizing the words: Tokenization is an important step before feeding the word embedding to the BERT model. Word tokenization is the process of splitting a large text into words. Tokenizing the words is needed because each word needs to be captured and subject to further analysis like classifying the document and counting them for a particular sentiment. The BERT ppre-trainedmodel's tokenizer was used for this purpose.

4. Padding the token vectors: Padding is a special form of masking where the masked token is at the start or the end of a sequence. Padding makes all sequences in a batch fit a given standard length. The BERT model's padding schemes are used to pad the corpus.

5. Masking the tokens: Masking is a way to tell sequence-processing layers that some inputs are missing; this helps to train the model better. We have applied masking with the BERT model.

6. Embedding with the BERT model: The final step before applying the classification algorithms is to get the BERT's contextual word embeddings. The numeric representation of documents with appropriate padding and masking is provided to the BERT model to produce powerful contextual embeddings of documents.

3.3 Design and Implementation

We can see the use case diagram for both registered and unregistered users or workflow diagram for our prototype in figure 11.

The use case for registered users

The Framework collects many documents by web scraping. We will be Classifying and Clustering them using our module. Then we place the documents in our document repository in a cloud-based service. The system creates a user profile as soon as a user registers in our system by gathering basic information (i.e., name, email, password, preference). Two tasks are performed parallelly when a user uploads a document for finding more similar ones, and the Framework updates the user profile and document retrieval. The document retrieval process matches the query to find the right domain and fetch the documents from our document repository. The system will also filter out records based on user preference from the user profile. Finally, some selected papers will appear to the user. A continuing monitoring system will be active the whole time to keep updating the user profile based on the search record. The recommender system will grow stronger with time as it learns more and more about user preferences.

The use case for unregistered users

The Framework also provides services for non-registered users. Users can enjoy the documents uploading and fetching similar records within the Framework. The Framework will not create any user profile; therefore, no filtering of the papers is performed.

3.3 Design and Implementation

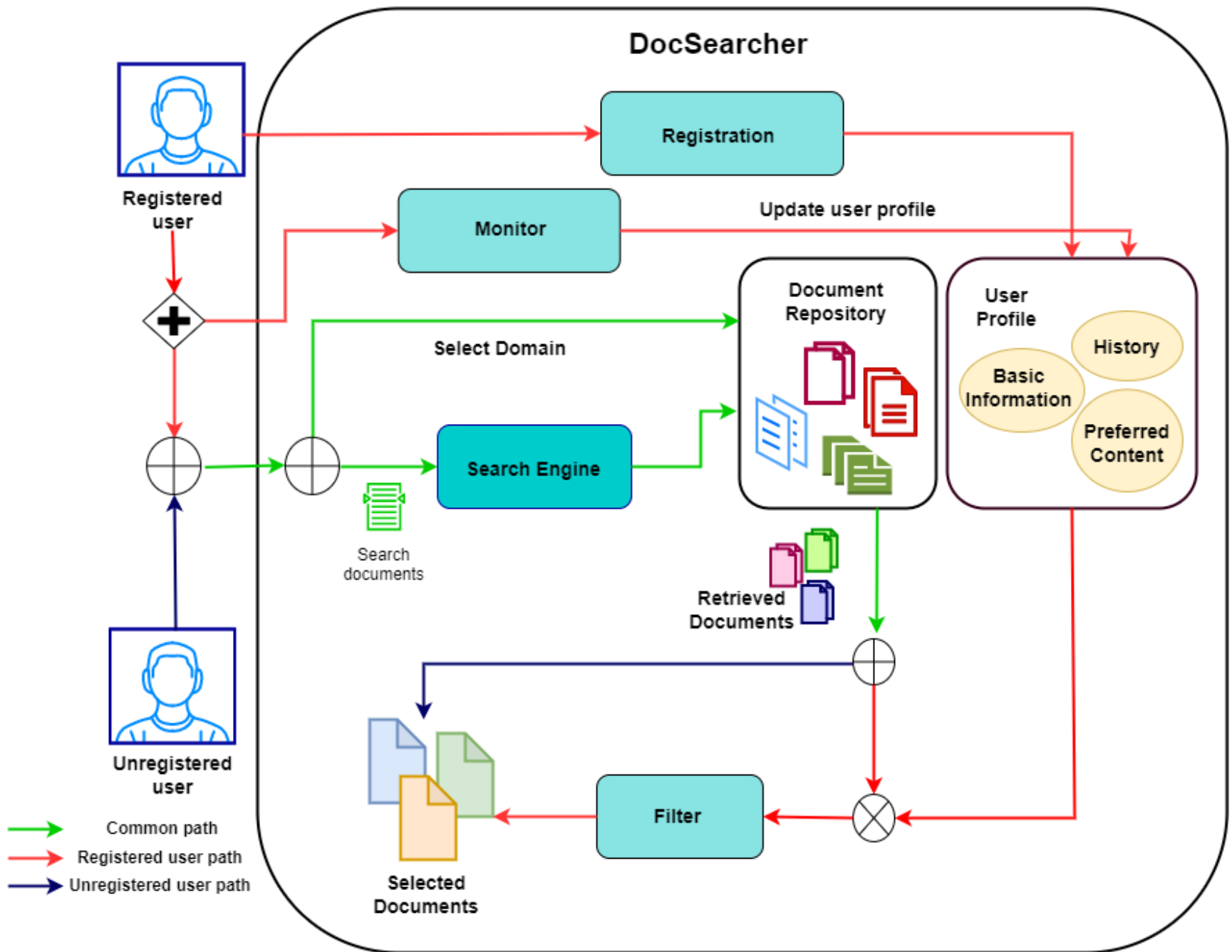


Figure 11: Workflow Diagram of the Framework

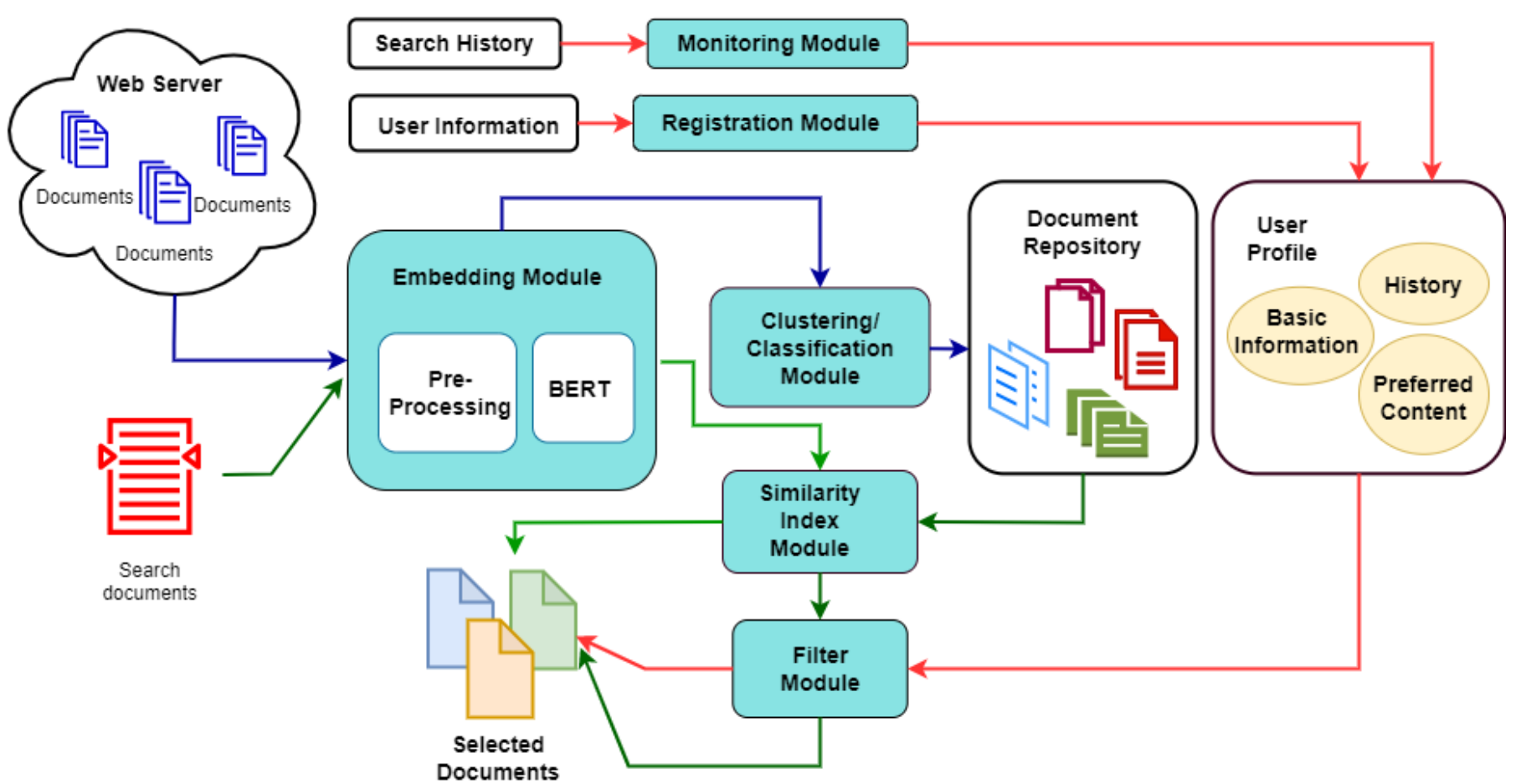


Figure 12: Prototype Framework

3.3 Design & Implementation

The framework (figure: 12) consists of several modules as follows:

Embeddings Module

In the embedding module, the developed system performs two significant tasks –pre-processing the text document and generating BERT embedding. The pre-trained language representation model BERT will generate contextualized sentence embeddings. Generally, this module maps each variable-length sentence in text documents to a 768-dimensional fixed-length sentence embedding. Two major components are connected to the embedding module, the Clustering/classification module, and the Similarity index module.

Classification and Clustering Module

BERT models' embeddings are then used to perform clustering and classification. As shown in figure 13, the clustering and classification methodologies consist of three modules, the embedding, classification and the clustering modules. The document vectors are used to apply classification and clustering algorithms.

Both distance-based and density-based clustering has been used to compare the results of different clustering methods. The Dunn index and silhouette coefficient were used for distance-based clustering and density-based clustering, and the silhouette coefficient was employed as the performance assessment matrices for clustering.

Dunn Index, The Dunn index attempts to assess the clustering's compactness and variance. If the variation between cluster members is modest, the cluster is termed compact. If the clusters are widely apart, they are termed well separated. A higher Dunn Index will indicate compact, well-separated clusters, while a lower index will indicate less compact or well-separated clusters.

$$Dunn\ index = \frac{\min(\text{inter cluster distance})}{\max(\text{intra cluster distance})}$$

Silhouette Coefficient, In comparison to other clusters, the silhouette value is a measure of how similar an object is to its cluster. A high number implies that the object is highly matched to its cluster and poorly compared to Silhouette surrounding clusters. The clustering setup is useful if the majority of the items have a high value. The clustering setup may have too many or too few clusters if many points have a low or negative value. Any distance metric, such as the Euclidean distance or the Manhattan distance, may be used to determine the silhouette.

$$Silhouette\ Coefficient = \frac{b-a}{\max(a,b)}$$

Here, a = mean intra – cluster distance and

b = mean nearest cluster distance for each sample

Silhouette Coefficient is only defined if the number of labels is $2 \leq n_{\text{labels}} \leq n_{\text{samples}} - 1$.

3.3 Design & Implementation

After the algorithm's performance is measured after the clustering of the embedding produced by BERT, as clustering is an unsupervised algorithm and obtaining the accuracy is often difficult, we employed a mapping method to evaluate the clustering performance. The cluster labels are mapped with the actual data labels. This is achieved by finding the Jaccard similarity of each cluster label with each domain label. The maximum of the Jaccard similarity index is mapped to the most similar domain. We can determine which cluster label is more identical to which domain in this way. The same method is then done in reverse to determine which domain is closest to which cluster. The Jaccard similarity measure yields a two-dimensional matrix. The largest value of each row determined the most common values in the tuples. We generate a list of tuples comprising the most common predicted cluster labels and actual data labels using the maximum values and indices. Finally, the process is repeated until each cluster label can be assigned to a domain label.

Jaccard Similarity, The size of the intersection divided by the size of the union of two sets yields the Jaccard similarity. The intersection and unions of two sets can be stated in set notation.

$$J = \frac{A \cap B}{A \cup B}$$

We have applied Decision Tree, Random Forest, and KNN (K-Nearest neighbors) algorithms for classification. We have adopted a stratified 10-fold cross-validation technique to achieve optimal accuracy. Finally, the documents are separated based on their context and stored in a cloud repository of the system.

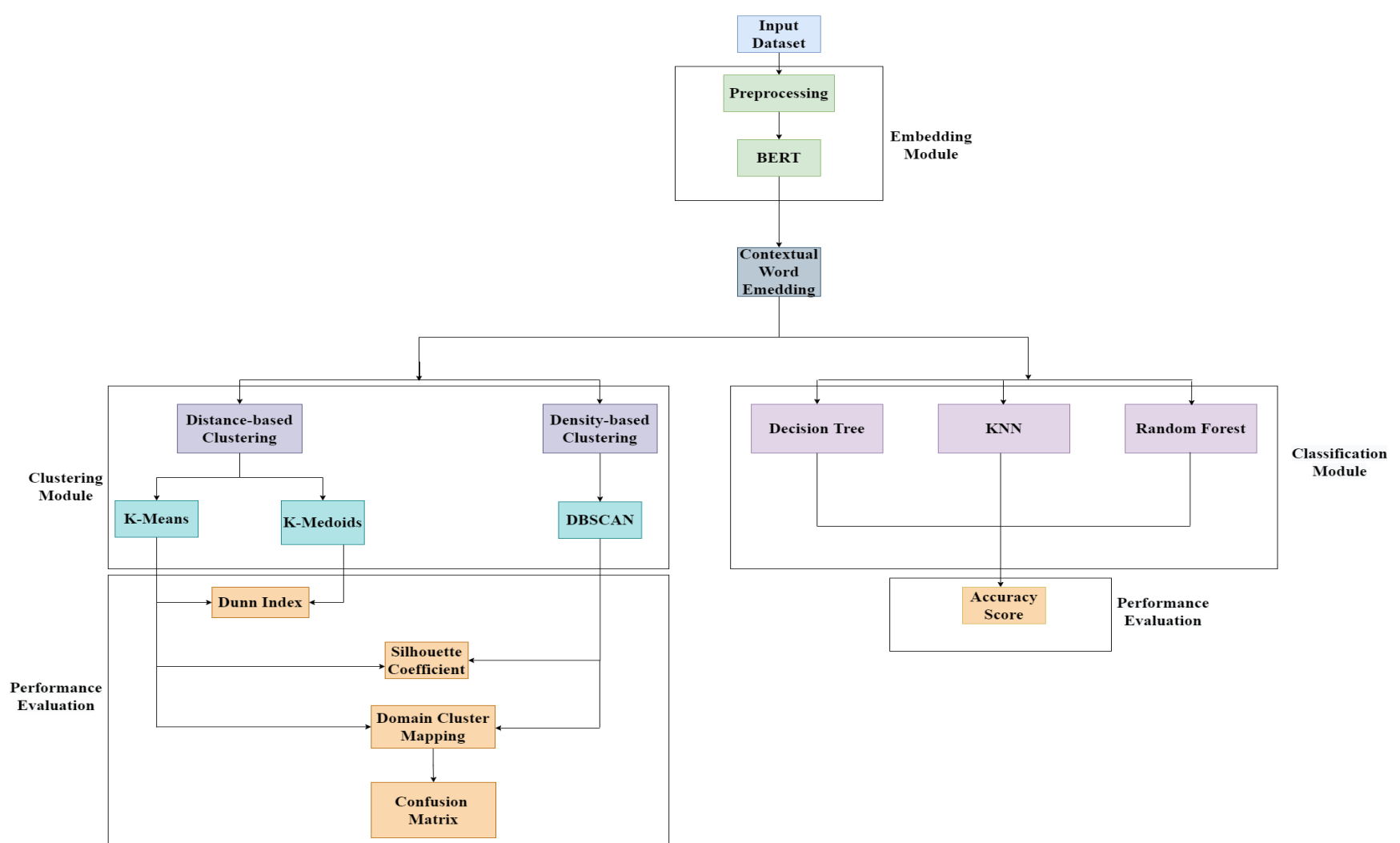


Figure 13: Classification and Clustering Framework

3.3 Design & Implementation

Algorithm 1 Cluster Domain Mapping Pseudocode

```
for  $i = 0$  to  $domainLength$  do
  for  $j = 0$  to  $domainLength$  do
     $JaccardSimClusterToDomain(i, j) \leftarrow JaccardSim(cluster[i], domain[j])$ 
     $JaccardSimDomainToCluster(i, j) \leftarrow JaccardSim(domain[i], cluster[j])$ 
  end for
end for

for  $i = 0$  to  $domainLength$  do
   $ClusterDomain.append(index[\max(JaccardSimClusterToDomain[i])])$ 
   $DomainCluster.append(index[\max(JaccardSimDomainToCluster[i])])$ 
end for

for  $i = 0$  to  $domainLength$  do
   $Similarity.append(JaccardSim(ClusterDomain[i], DomainCluster[j]))$ 
end for

for  $i = 0$  to  $domainLength$  do
  if  $similarity[i] > 0$  then
     $mapping.append(ClusterDomain[i] \cap DomainCluster[i])$ 
  else
     $mapping.append(-1)$ 
  end if
end for

while  $-1$  in  $mapping$  do

  if  $count(-1)$  in  $mapping = 1$  then
     $cluster \leftarrow find(index(mapping))$ 
     $domain \leftarrow find(domains) \text{ not in } mapping$ 
  end if

   $mapping[cluster] \leftarrow domain$ 
   $remainingCluster \leftarrow [find(index) \text{ where } mapping = -1]$ 
   $remainingDomain \leftarrow [find(domain) \text{ not in } mapping]$ 

  while  $i < length(remainingCluster)$  do
    while  $j < length(remainingDomain)$  do
       $matrix.append(JaccardSimClusterToDomain[remainingCluster[i]], [remainingDomain[j]])$ 
       $i ++, j ++$ 
    end while
  end while

  for  $i = 0$  to  $domainLength$  do
     $ClusterDomain2.append(index[\max(matrix[i])])$ 
     $DomainCluster2.append(index[\max(matrix.T[i])])$ 
  end for

  for  $i = 0$  to  $domainLength$  do
     $Similarity2.append(JaccardSim(ClusterDomain2[i], DomainCluster2[i]))$ 
  end for

  for  $i = 0$  to  $domainLength$  do
    if  $similarity[i] > 0$  then
       $mapping.append(ClusterDomain[i] \cap (DomainCluster[i]))$ 
    end if
  end for

end while
```

Registration Module

For authentication and user profile creation, a registration module will be implemented. It will gather some basic information of the user to create a profile.

Monitoring Module

The monitoring module is created based on the user profile and preference. The user searches for content will stay in the user profile model. With time the model updates by adding more preferred items and discarding redundant items. The algorithm can produce statistics on the number of times a domain has been explored and update the user preference making the user-centric search **system** more relevant and updated.

Filter / Recommender Module

The recommender module will perform in the following manner, at first, documents are fetched based on the user search query. Secondly, the records are filtered out based on user preferences. This filter module will communicate with the user profile model to collect the preferences.

Similarity index Module

The similarity index module retrieves the most similar documents per the user query. This module works as an alternative to the clustering module. Users can find similar documents by uploading one into the system if they want. Our Framework offers both domain-dependent and domain-independent similar document retrieval options.

Based on the similarity index, documents are shown to the user. The similarity index module retrieves the most similar documents. We have used the cosine similarity measures, and the framework retrieves the five most similar documents. The definition of cosine-based similarity is mentioned below, in which A and B are vectors representing two documents.

$$\text{Cosine Similarity}(A, B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n A_i^2} \times \sqrt{\sum_{i=1}^n B_i^2}}$$

3.4 Project management Features

This section has explained the project management process of the capstone project. The Prototype was developed with all the project management concepts to reflect a real-life project development. This capstone prototype was built to expand our understanding of deployment-level project management. Developing required documentation for a project has helped us gain essential knowledge about the capstone project's budgeting, resource allocation, and work breakdown structure. Developing this capstone project from an entrepreneurial point of view has enlarged our understanding of focusing on the work that matters, free from the distractions caused by tasks going off track or budgets spinning out of control. Considering it as a professional project. It has required around 4-5 months to build this web-based application except for the Research and literature study part.

3.4.1 Business Model Canvas

Figure 14 shows the prototype's BMC (Business Model Canvas). The BMC is an integral part of project management. BMC consists of the nine essential building blocks explaining crucial information about the project as a business area. Identifying planning gaps between these nine building blocks is the actual value of the Business Model Canvas. It is an all-in-one canvas that explains what the prototype offers, features, the target customers, the value the prototype provides, how we as a provider make relationships with the customers, the key strategy of the model, and the cost-revenue structure. For the capstone prototype, we have developed the BMC showing the value proposition and explaining what unique features we are offering. We also have identified the channels to stay connected to the customers, our key business partner in the development environment.

3.4 Project Management Features

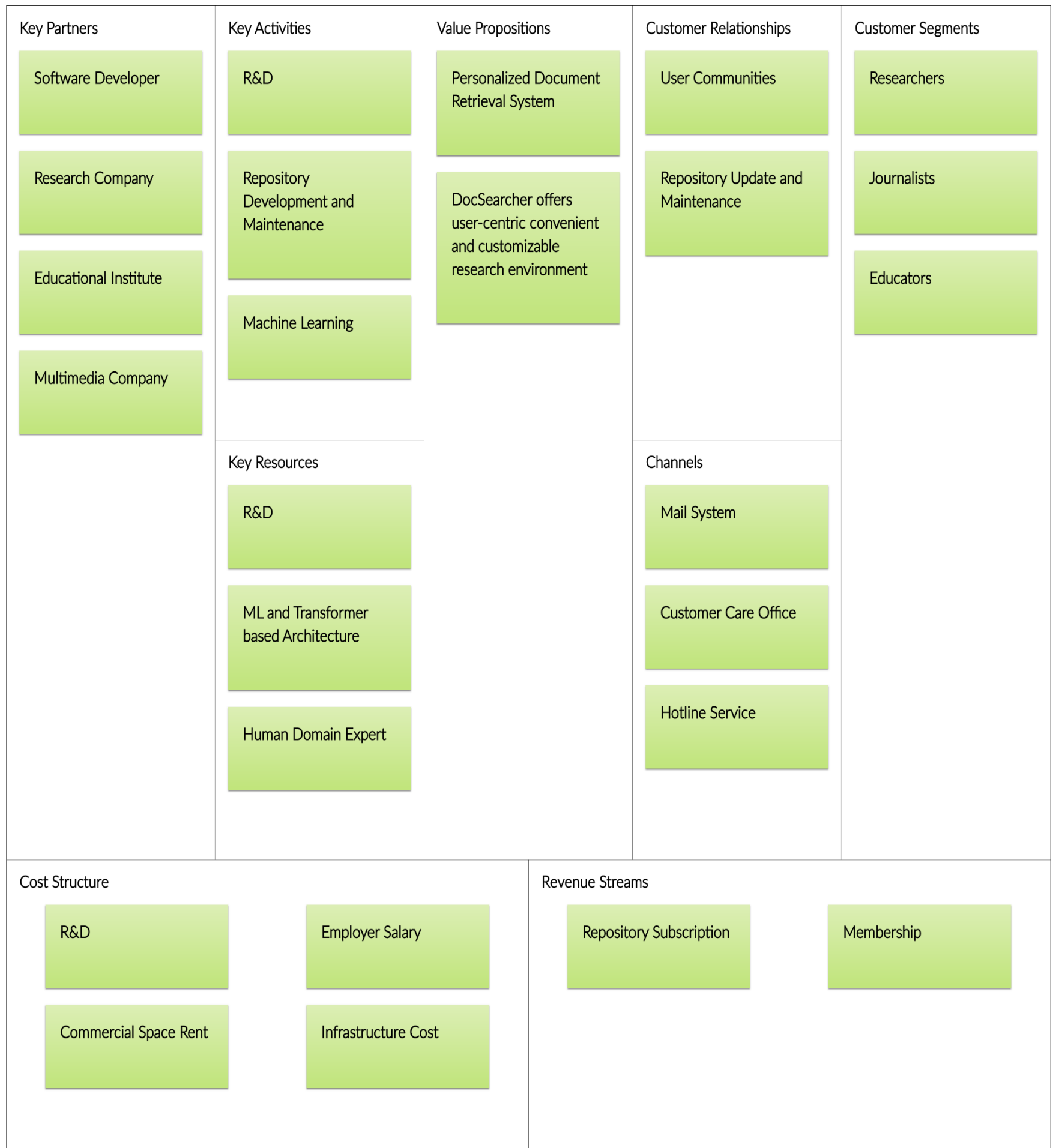


Figure 14: Business Model Canvas

3.4.2 Work Break Down Structure

A work breakdown structure (WBS) is a visual, hierarchical, and deliverable-oriented deconstruction of a project. It is a helpful diagram for project managers because it allows them to break down their project scope and visualize all the tasks required to complete their projects [26]. It is a tool that can be used for projects, programs, and even initiatives to understand the work that has to be done to successfully produce a deliverable [27].

3.4 Project Management Features

The WBS of our project shows in figure 15 all the work required to complete this project. It provides a visual of the entire scope of the project. This is also used for allocating which employee should be assigned for which work. By calculating the cost per task, the total budget of the product can be estimated. Also, this dividing task into multiple sections is constructive for resource allocation. Lastly, this helps the team feel invested in the planning. The WBS was developed with the original budget of 8,76,233 BDT. The WBS for the project was divided into five main tasks, namely the design phase, the dataset collection phase, the implementation phase, the testing phase, and finally, the maintenance phase of the prototype.

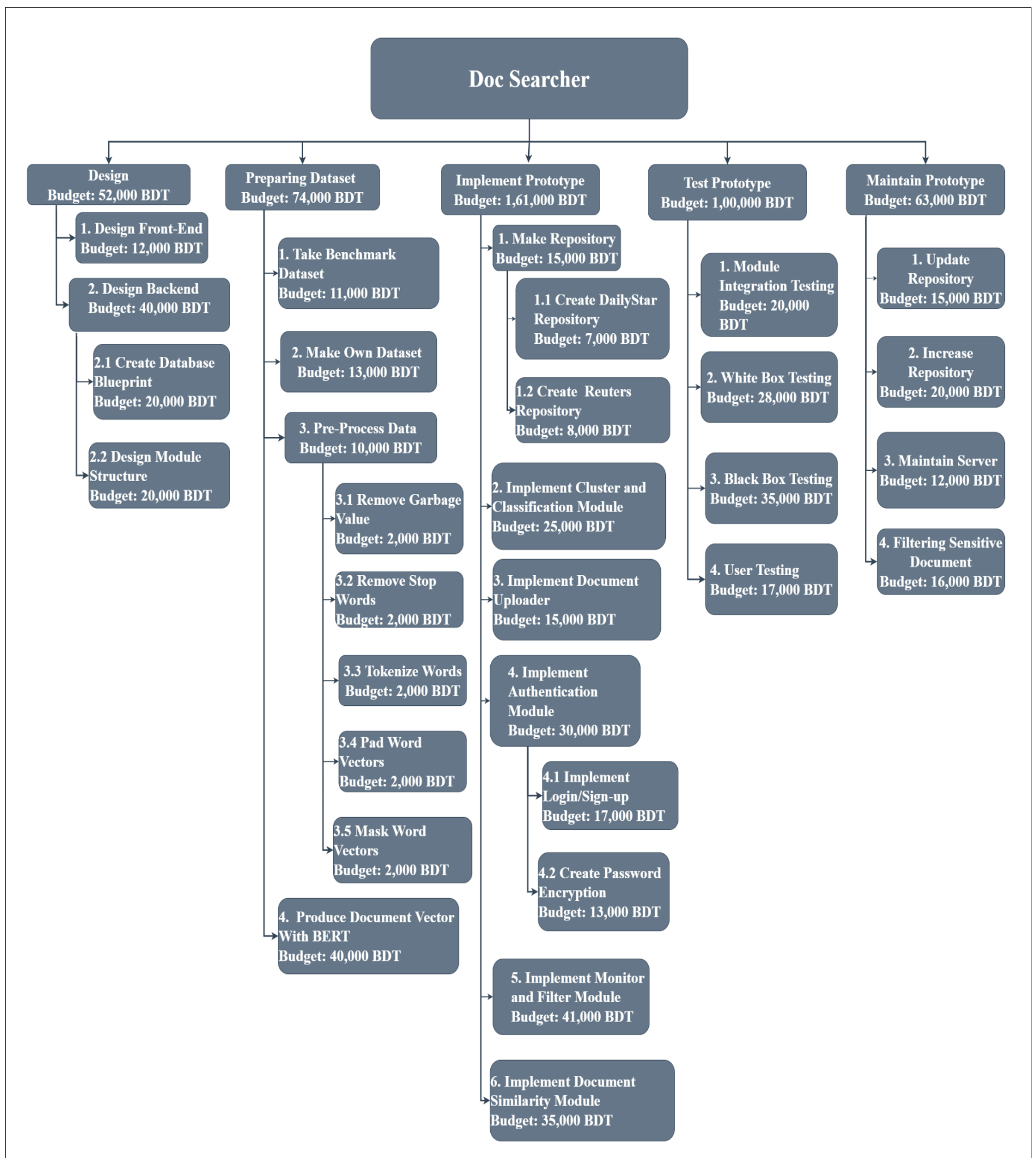


Figure 15: Work Breakdown Structure (WBS) of DocSearcher

3.4 Project Management Features

3.4.3 Resource Allocation

Resource allocation is strategically selecting and assigning available resources to a task or project to support business objectives. In the context of accounting, resource allocation deals with setting people and their skills to projects, also known as engagements [28]. For our project, we have designated days for each task as per the importance and time required for the task shown in table 4. Each work is given to the respective employee. For allocating the task's budget, we've calculated their salary, resource cost, and how long it takes to fulfill that task. The total cost for this project, including development, infrastructure, maintenance, and employee salary, will be 8,76,233 BDT. Table 5 shows the budget for employee person-month salary for our development part of the project.

In our project, four roles are included such as- project manager, front-end developer, back-end developer, and quality assurance officer.

Table 4: Resource Allocation

Third Level Breakdown	Responsible	Allocated Days	Start Time	End Time	Predecessors
1. Design front-end	Front end Developer	9d	10/15/21	10/27/21	---
2. Design module structure	Back-end Developer	7d	10/20/21	10/28/21	---
3. Design database structure	Back-end Developer	5d	10/20/21	10/26/21	---
4. Take Benchmark Dataset	Back-end Developer	5d	10/25/21	10/29/21	---
5. Scrap Data from DailyStar.net	Back-end Developer	7d	10/30/21	11/08/21	---
6. Remove Garbage Value	Back-end Developer	7d	11/09/21	11/17/21	4, 5
7. Remove Stop Words	Back-end Developer	1d	11/18/21	11/18/21	6
8. Tokenize Words	Back-end Developer	1d	11/19/21	11/19/21	7
9. Pad Word Vectors	Back-end Developer	1d	11/22/21	11/22/21	8
10. Mask Word Vectors	Back-end Developer	1d	11/23/21	11/23/21	9
11. Produce Document Vector With BERT	Back-end Developer	16d	11/24/21	12/15/21	10
12. Create Daily Star Repository	Back-end Developer	6d	12/16/21	12/23/21	1, 3, 11
13. Create Reuters Repository	Back-end Developer	7d	12/16/21	12/24/21	1, 3, 11

3.4 Project Management Features

Third Level Breakdown	Responsible	Allocated Days	Start Time	End Time	Predecessors
14. Implement Cluster and Classification Module	Back-end Developer	12d	12/27/21	01/11/22	2, 12, 13
15. Implement Document Uploader	Back-end Developer	5d	01/12/22	01/18/22	14
16. Implement Login/Sign-up	Back-end Developer	8d	10/29/21	11/09/21	1, 2, 3
17. Create Password Encryption	Back-end Developer	4d	11/10/21	11/15/21	16
18. Implement Monitor and Filter Module	Back-end Developer	7d	01/19/22	01/27/22	15, 17
19. Implement Document Similarity Module	Back-end Developer	12d	01/28/22	02/14/22	18
20. Module Integration Testing	Quality Assurance Engineer	8d	02/18/22	03/01/22	20
21. Design Testing Structure	Quality Assurance Engineer	3d	02/15/22	02/17/22	19
22. White Box Testing	Quality Assurance Engineer	7d	02/18/22	02/28/22	20
23. Black Box Testing	Quality Assurance Engineer	11d	02/18/22	03/04/22	20
24. User Testing	Quality Assurance Engineer	10d	03/07/22	03/18/22	21, 22, 23

Table 5 : Employee's Salary Estimation

Role	Salary per month
Front End Developer	40,000 BDT
Back End Developer	70,000 BDT
Quality Assurance Engineer	1,00,000 BDT

3.4.4 Critical Path Method

Critical Path Method (CPM) is an algorithm for planning, managing, and analyzing the timing of a project. The step-by-step CPM system helps identify critical and non-critical tasks from projects start to completion and prevents temporary risks [29]. CPM helps calculate the time and resources needed for the project and find the critical paths for the project so there's no scheduling problem. In critical paths, if a task is finished on a particular day, another work will start on the same day. So any kind of lapse in these paths will delay the entire project. From figure 16, we can conclude that the estimated critical duration is 102 days.

CPM : DocSearcher

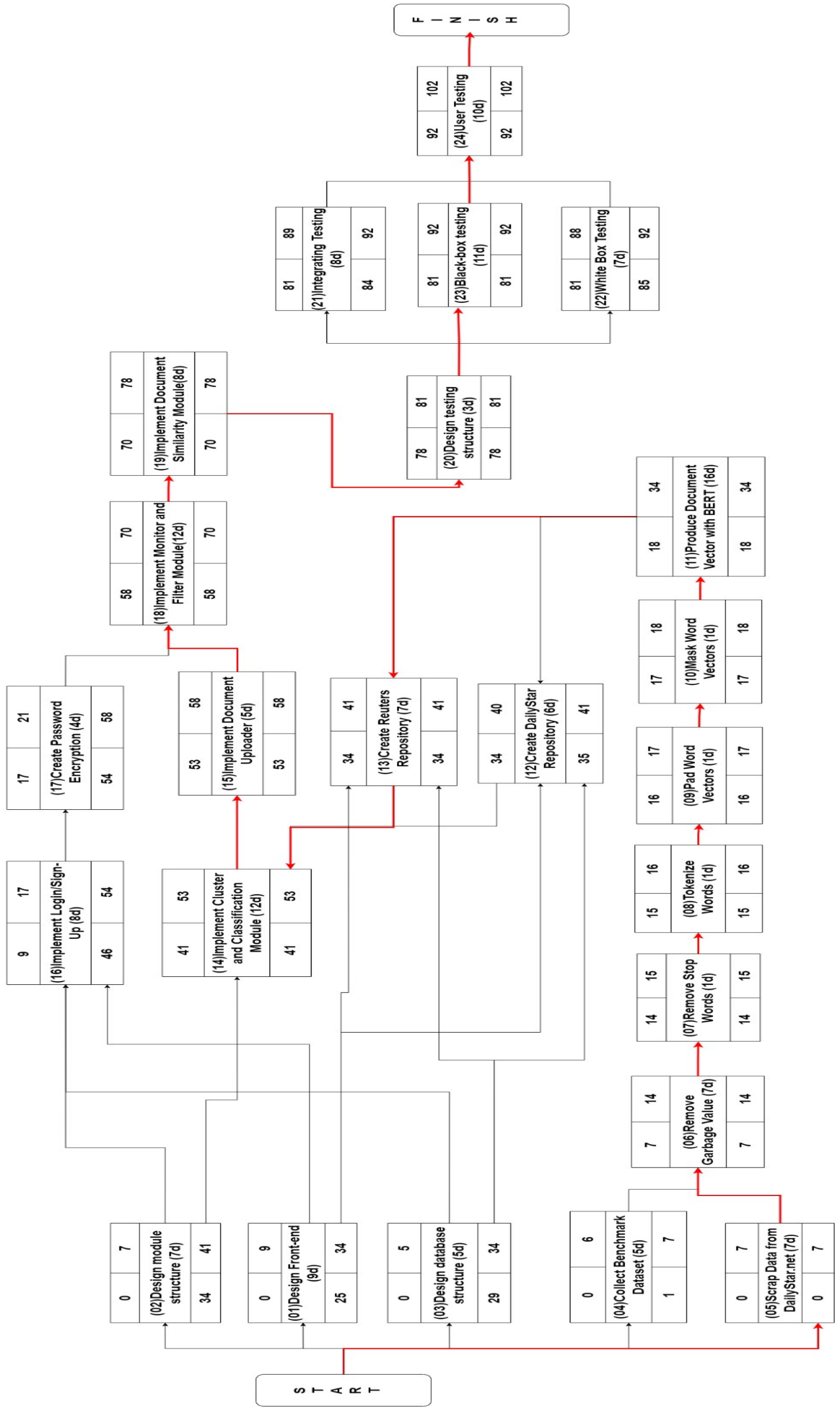


Figure 16: Critical Path Method (CPM)

3.4 Project Management Features

3.4.5 Infrastructure Cost

To design, develop, implement, and expand the infrastructure there are certain costs that are needed. The detailed infrastructure costing for our capstone project is given in table 6.

Table 6: Infrastructure Cost

Infrastructure Cost				
Serial	Equipment	Description	Quantity	Cost
1	Laptop/PC	Intel(R) Core (TM) i5-8250U CPU 1.60GHz 1.80 GHz. RAM 16.0 GB. x64-based processor. Windows operating system	3 Laptop/PC (min)	2,10,000 BDT
2	Google Colab Pro	RAM: 53GB Number of CPU: 2 Disk Space: 150 GB	1 Account (min)	23,233 BDT
4	TP-Link Router	TP-Link Wireless N Router WR841N Model No. TL- WR841N Mode: 11bgn mixed Internet: 100Mbps full duplex SPI FireWall: enabled	1 Piece	4,000 BDT
Total				2,37,233 BDT

Budget = 4,50,000 BDT, Infrastructure Cost = 2,73,233 BDT

Total = 7,23,233 BDT

3.4.6 Break-Even Point Calculation

A break-even point means a point at which the total cost and revenue will be the same in a business. In other words, a point in production where the expenses of production are equal to the product's revenues. It is used to calculate the number of units needed to cover the total cost for the end product.

Assuming,

At initial stage, Customer requirement = 2,

Account requirement per month = 30

Package 1: 1 customized repository and 10 user account, Cost = 50,000 BDT

Additional Per Account Adding Cost = 1,000 BDT

If package 1 is sold to two customers, then revenue will be = 1,00,000 BDT

Account left = 30 – 20 = 10, Other 10 accounts revenue = 10,000 BDT

Per month revenue = 1,10,000 BDT

As per the projection after around 7 months there will be a break-even point. So, this project can expect to get revenue after 7 months.

3.5 Test Structure

3.5 Test Structure

We used four test scenarios to put our prototype to the test shown in table 7. We began by testing the system's sign-up mechanism. It displays the result in accordance with the rules for the input of test data. The login function is the following function we tested; there were four test cases for it. If a user can log in successfully, this function has been correctly implemented. If the details of the repositories, such as repository pages, and domain/topic list, are displayed for the test repository pages function, then it is routing correctly. Finally, when testing the document uploader, the result will show appropriate classification if a user uploads a text file after selecting a repository.

Table 7: Test Structure

Project Name				Docsearcher		
Module Names				Sign up module, Login Module, Document upload module.		
Created By				Nazia, Sumona, Kashfia		
Creation Date				May 13, 2022		
Test Scenario ID	Test Scenario	Test Case ID	Test Case Description	Test Data	Conditions	Expected result
TS-01	Test Signup Functionality	S-1	1)Enter a valid email 2)Enter a valid password 3)Click login	1)Alice@gmail.com 2)alice1234	See Login Page	Successful Sign Up
		S-2	1)Enter invalid email 2)Enter valid password 3)Click login	1)abc.com 2)alice1234	Redirect to Sign Up page	Unsuccessful Sign Up
		S-3	1)Enter valid email 2)Enter invalid password 3)Click login	1)Alice@gmail.com 2)1234567	Redirect to Sign Up page	Unsuccessful Sign Up
		S-4	1)Enter invalid email 2)Enter invalid password 3)Click login	1)abc.com 2)123456	Redirect to Sign Up page	Unsuccessful Sign Up
TS-02	Test Login Functionality	L-1	1)Enter valid email 2)Enter valid password 3)Click login	1)Alice@gmail.com 2)alice1234	See Home Page	Successful Login

3.5 Test Structure

Test Scenario ID	Test Scenario	Test Case ID	Test Case Description	Test Data	Conditions	Expected result
		L-2	1)Enter an invalid email 2)Enter valid password 3)Click login	1)abc.com 2)alice1234	Redirect to login page	Unsuccessful Login
		L-3	1)Enter a valid email 2)Enter invalid password 3)Click login	1)Alice@gmail.com 2)1234567	Redirect to login page	Unsuccessful Login
		L-4	1)Enter invalid email 2)Enter invalid password 3)Click login	1)abc.com 2)123456	Redirect to login page	Unsuccessful Login
TS-03	Test Repository pages	R-1	1)Click home Page 2)Click any repository 3)Click Domain/Topics	HTTP request	1)Viewing repository page 2) Viewing domain list/topic list page	Successful routing
TS-04	Testing Document Uploader	D-1	1)Choose a repository 2) Choose a .txt file 3) Click submit	1) Daily Star/Reuters 2) Testfile1.txt	Redirecting to documents page with suggested domains and similar documents.	Successful classification
		D-2	1)Choose a repository 2)Input an invalid file format 3) Click submit	1) Daily Star/Reuters 2) File.pdf	Redirecting to uploader	Un-Successful classification
		D-2	1)Choose a repository 2)Input an empty file 3) Click submit	1) Daily Star/Reuters 2) empty_file.txt	Redirecting to uploader	Un-Successful classification
		D-2	1)Choose a repository 2)Input an invalid file (not English) 3) Click submit	1) Daily star/Reuters 2) invalid file	Redirecting to uploader	Un-Successful classification

3.6 Materials and Devices

This section briefly described the functionality of the modern engineering and machine learning tools we used to research and develop for Capstone study.

Operating System and Hardware Configuration

In this capstone project, we have conducted all the experiments in a Windows operating system. The system configuration used in this project; Intel(R) Core(TM) i5-8250U CPU 1.60GHz 1.80 GHz Installed RAM 16.0 GB. The system type is a 64-bit operating system, an x64-based processor with Windows 11 OS.

Python Machine Learning Libraries

Hugging Face: We have used the hugging face model BERT to develop the prototype and research purposes. Hugging Face is a startup in the Natural Language Processing (NLP) domain, offering many modern state-of-art models, including the Transformer-based models.

Python's SciKit-Learn: Python's SciKit-Learn is a simple and efficient library for predictive data analysis and implementing Machine Learning algorithms. We have used the offered libraries for data analysis, pre-processing, cleaning, and analyzing the machine learning outcomes during our research.

Development Framework

Google Collab is a powerful python coding tool offering Zero configuration, Free accessed GPUs, and Easy sharing. Collab is a platform provided by Google to write codes in python with all the libraries. The platform offers more GPU power to run heavy models. We have used Google's collab services to conduct the research and analysis results for the capstone project.

Python Flask for Application Prototyping

The Flask is a microframework developed in python. The Framework is a modern and popular tool that builds machine learning models and hosts them in a cloud-based server. We have used this modern tool to construct the proposed prototype in this capstone project.

Application Integrated development environment (IDE)

We have used Visual studio code IDE to develop the proposed prototype. Visual Studio Code is a source-code editor. VS code is a modern popular editor that supports debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git.

GitHub for version control

GitHub is a popular safe version control system. We have used the functionality of GitHub to maintain version control and source code management.

Results

This chapter has explained the result analysis of the research. The clustering and classification result analysis has been shown in this chapter to evaluate the performances of the algorithms. The results chapter of the research report will endeavor to tell the findings without attempting to analyze or assess them, as well as give guidance to the discussion section. The findings are presented, and the analysis is revealed.

4.1 Classification of Reuter Dataset

4.1 Classification of Reuter Dataset

The following section presents the result analysis for the Reuters and Daily Star Dataset with the proposed methods. We have applied the classification algorithms to analyze which classifier works best and the influence of the dataset size in the classification algorithms. Classifications accuracy is much higher with the increasing length of the dataset. We have applied stratified k-fold as a validation technique, for analysis, we have taken K=10, and the stratified k-fold cross-validation gives us the min-max and average accuracy of the classifier and the standard deviation. The cross-validation helped us gain optimal results for classification algorithms applied in this capstone project. The overall classifier algorithm performance is shown in table 8. The random forest classifier outperforms than the others by all the datasets on average.

Table 8: Classifier Algorithm performance analysis on Reuter Dataset

Dataset Name	Data size	Decision Tree				Random Forest				KNN			
		Max (%)	Min (%)	Avg (%)	std	Max (%)	Min (%)	Avg (%)	std	Max (%)	Min (%)	Avg (%)	std
DS1	200	85	65	76.5	0.067	95	75	88.0	0.059	90	75	81.5	0.047
DS2	500	84	66	75.8	0.059	90	84	87.0	0.022	90	78	85.6	0.047
DS3	1000	76	63	67.7	0.035	94	83	88.0	0.039	92	81	88.5	0.033
DS4	5000	75	70	72.8	0.014	92	88	90.0	0.015	93	87	89.8	0.017

The following figures (fig. 17 and fig. 18) present a visualized comparison of the performance of the classifier algorithm on the Reuter Dataset.

4.1 Classification of Reuter Dataset

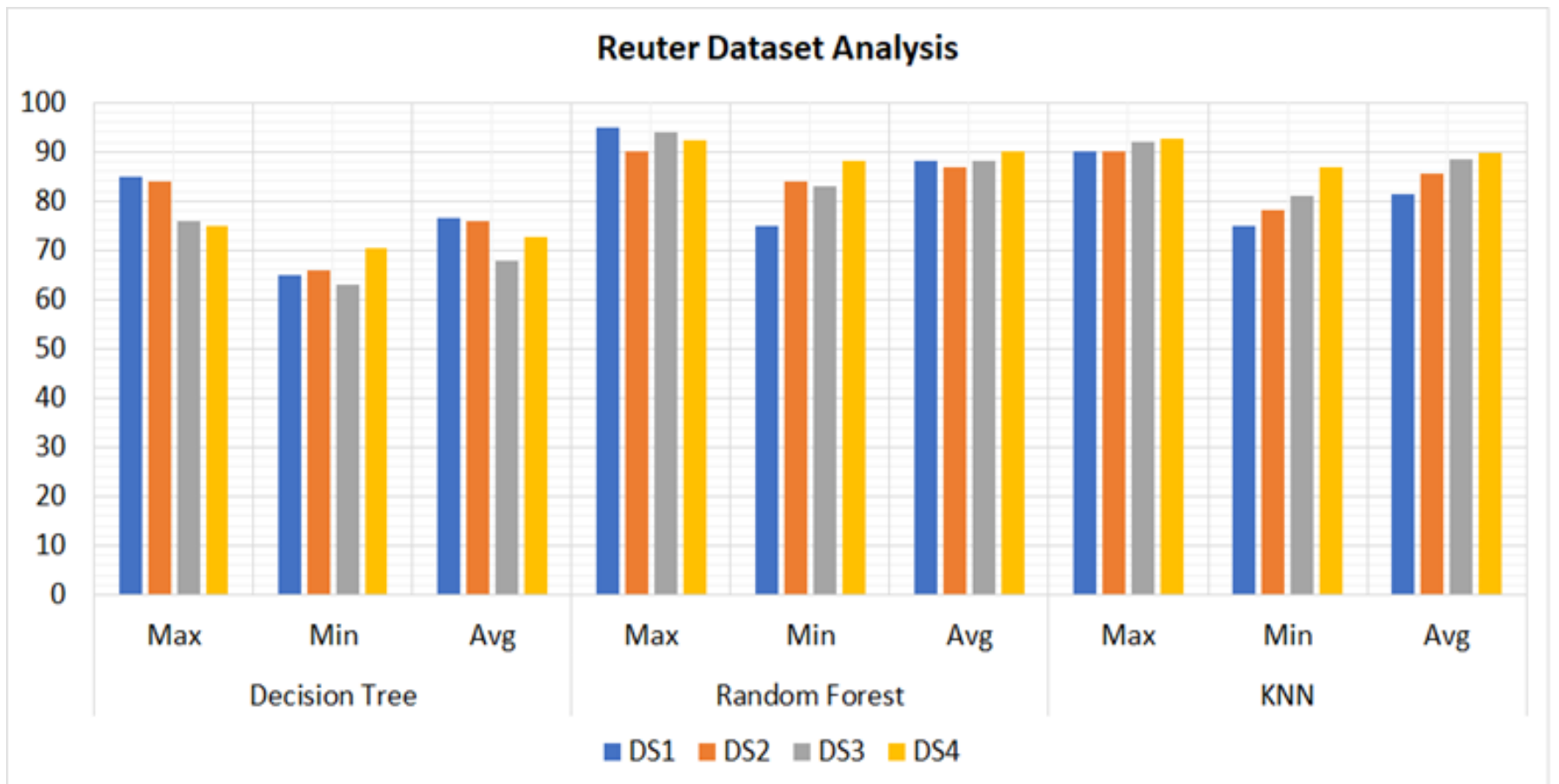


Figure 17 : Reuter Dataset Analysis

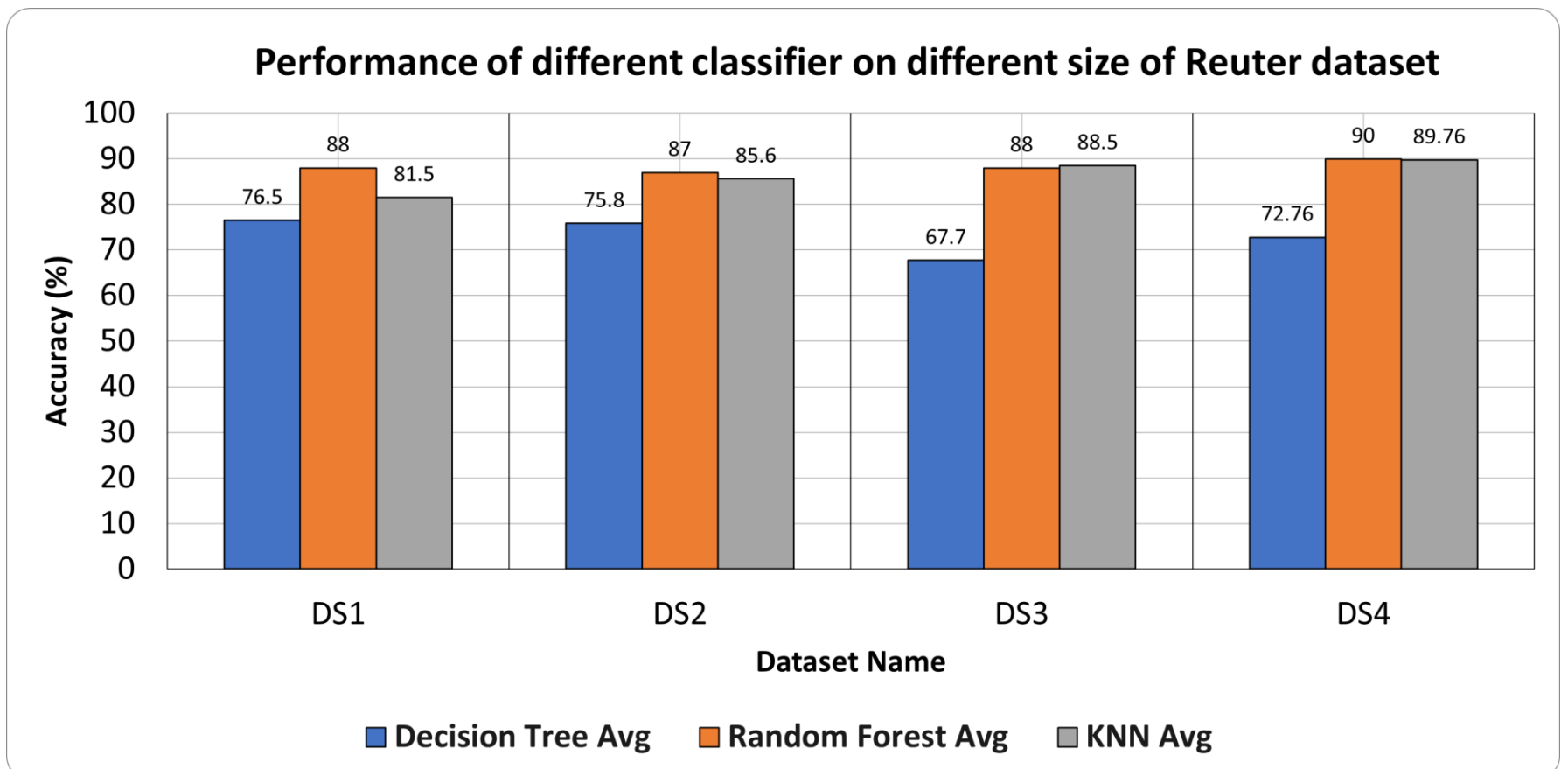


Figure 18: Different Classifiers on Reuter Dataset Size vs. Accuracy

Performance Evaluation on Reuter Dataset

In this section, we have compared the performance and accuracy of our research to the research work of Yutong Li et al. [17] weighted BERT model. The proposed methods in this paper achieve higher classification accuracy than Yutong Li et al. [17]. The proposed algorithms are simple yet effective. The classification algorithms obtained in this capstone project outperform the accuracy in Yutong Li et al. [17].

4.1 Classification of Reuter Dataset

Table 9: Performace Comparison with the target research

Dataset name	Dataset Size	Topic number	Decision Tree Accuracy	Random Forest Accuracy	KNN Accuracy	Target Accuracy
DS4_v2	200	4	76.0	91.5	91.0	75.0
DS5_V2	500	5	74.4	86.4	85.2	63.0
DS8_v2	1000	8	66.2	82.6	81.5	53.7
DS15_v2	5000	15	59.8	84.7	85.1	67.9

Figure 19 is our research work's performance analysis visual representation with the target study mentioned in Yutong Li et al. [17]. We have taken this particular research to compare the performance of the algorithms because we have also worked with Reuters Dataset similar to Yutong Li et al. [17]. The sectioning of the dataset was also kept identical to the target research to compare and evaluate the results. The proposed methodologies and algorithms achieved 86.4% accuracy on Random Forest Classifier, surpassing the highest accuracy obtained by the target research by 21.485%.

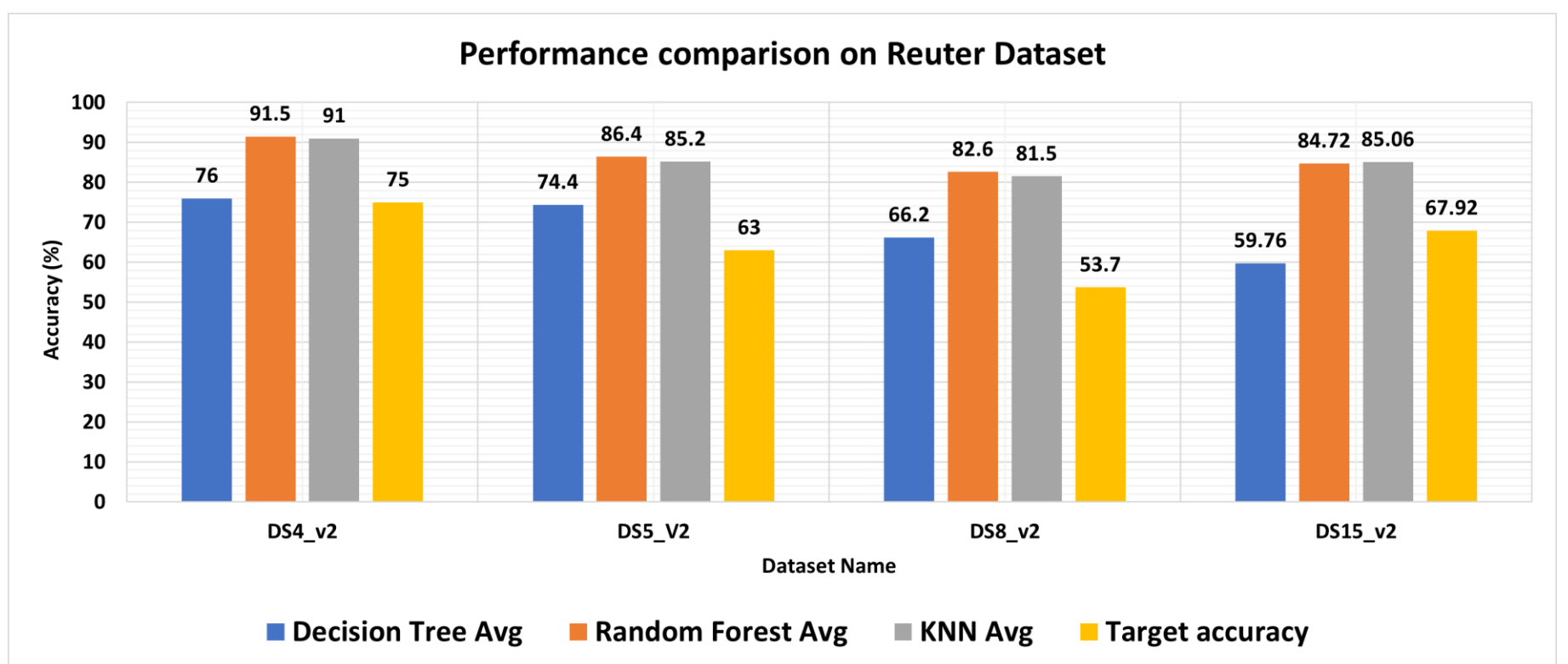


Figure 19: Performance comparison of Capstone's accuracy with the targeted research work on Reuters Dataset.

4.2 Classifier Analysis on Daily Star Dataset

We also applied a classifier algorithm on the "Daily Star" Dataset to analyze the classification algorithms for our capstone project. For this, we have divided the dataset into two sections, DS415 and DS212. The accuracy is obtained by the classifiers' stratified k-fold cross validation's min, max, and average accuracy scores and standard deviation. The random forest classifier performs best in both datasets, as in table 10.

Table 10: Classifier Algorithm performance analysis on Daily Star Dataset

Dataset Size	Decision tree(%)	Random Forest(%)	KNN(%)
DS415	36.61	75.92	62.86
DS212	31.17	66.48	60.47

In Figure 20, the visual representation is described of the classifier algorithm performance in the Daily Star dataset.

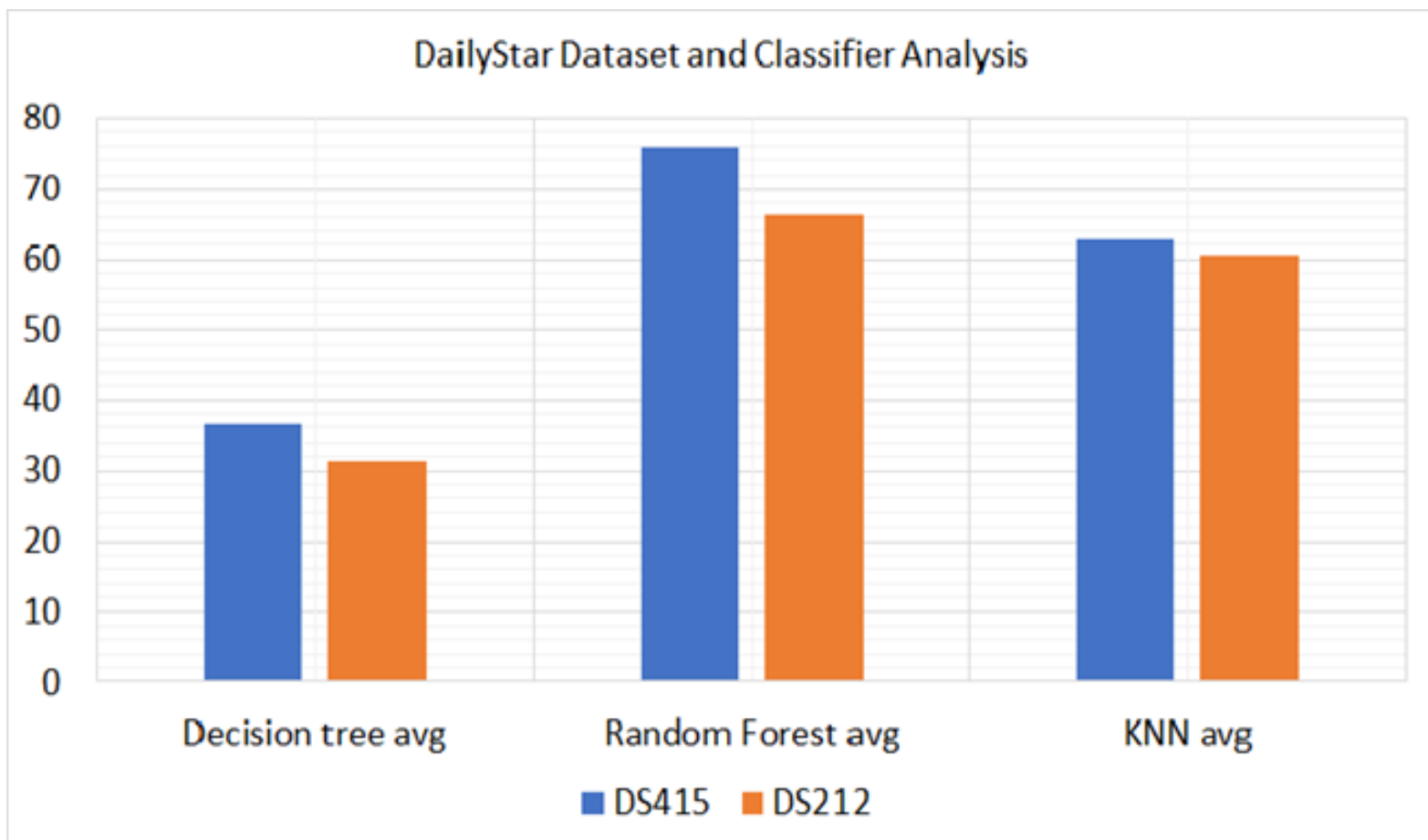


Figure 20: Daily Star Dataset and Classifier Analysis

Table 11: Test Data analysis of Daily Star Dataset

Sample No.	System Generated class label	Domain Expert Identified class label	Algorithm predicted class label		
			Decision Tree	Random Forest	KNN
1	environment	business	tech_news	business	environment
2	environment	environment	environment	business	environment
3	Sports	sports	sports	sports	sports
4	Sports	sports	tech_news	sports	sports
5	business	business	entertainment	business	business
6	business	tech_news	tech_news	business	environment
7	entertainment	entertainment	environment	business	sports
8	entertainment	entertainment	entertainment	entertainment	entertainment
9	tech_news	tech_news	tech_news	tech_news	tech_news
10	tech_news	business	tech_news	business	business
11	life_living	life_living	entertainment	life_living	life_living
12	life_living	life_living	environment	tech_news	tech_news
13	youth	tech_news	tech_news	entertainment	youth
14	youth	youth	entertainment	life_living	youth

Table 12: Unseen Test Data class label comparison between classifier predictions and System generated Vs. classifier predictions and Domain experts identified

	Accuracy		
	Decision Tree	Random Forest	KNN
System Generated class label	5/14	7/14	9/14
Domain Expert Identified class label	6/14	8/14	9/14

Performance Evaluation for Daily Star Dataset

Table 11 represents the test class label prediction containing 14 Daily Star news portal article samples as test data. We have taken help from a linguistic expert for a strong validation of our result. We have tested the Test dataset with a linguistic expert who can identify the domain by reading the Daily Star Dataset articles. As machine learning algorithms do not understand textual data, it would be adequate to validate the test dataset from a human domain expert.

4.3 Clustering on Daily Star Dataset

Performance of K-Means on Daily star Dataset

This section has explained the performance analysis of the clustering algorithms on the Daily Star dataset. The daily star dataset has seven domains. K-Means and K-Medoid method has been applied for clustering. Cluster numbers ranging from five to ten for finding optimal cluster numbers were taken. To assess the result, the silhouette score and Dunn index were calculated. Table 13 demonstrates that the Dunn index and Shallotte coefficient matrices produce the maximum accuracy for cluster number 5.

Table 13: KMeans and KMedoid Cluster Quality Analysis

cluster number	Dunn index K-Means	Silhouette avg K-Means	Dunn index K-Medoids	Silhouette avg K-Medoids
5	0.654415	0.135175	4.275027e-08	-0.032062
6	0.633541	0.128418	4.275292e-08	-0.049176
7	0.578599	0.082912	4.431572e-08	-0.030276
8	0.550345	0.081244	4.431572e-08	-0.021859
9	0.491421	0.074516	4.431572e-08	-0.026283
10	0.572035	0.084121	4.460831e-08	-0.019459

Performance of DBSCAN on Daily Star Dataset

Experiments on the Daily Star dataset were also conducted by performing feature reduction of the input dataset. The pre-trained BERT generated 768 embeddings which were reduced to 100 samples. Feature reduction was necessary as clustering algorithms tend to produce wrong predictions with large input vectors. UMAP was used to perform feature reduction. DBSCAN algorithm was applied afterward. For experimental purposes, the epsilon value ranging from 0.38 to 0.70 was taken for testing, shown in Table 14. The analysis of the DBSCAN algorithm revealed the fewest non-clustered documents, 27 were found for cluster number 7, and the corresponding silhouette score and epsilon values are 0.064687 and 0.64, respectively.

Table 14: DBSCAN Cluster Analysis

epsilon	Number of clusters	Number of Non clustered doc	Silhouette coefficient
0.38	2	385	-0.026521
0.40	3	367	-0.078938
0.42	6	314	-0.047469
0.44	8	271	-0.051195
0.46	8	223	-0.023049
0.48	11	155	0.005511
0.50	12	129	0.017165
0.52	11	97	0.030152
0.54	11	73	0.041800
0.56	10	50	0.070519
0.58	10	36	0.071821
0.60	8	32	0.067716
0.62	8	31	0.067656
0.64	7	27	0.064687
0.66	6	23	0.063110
0.68	5	22	0.062567
0.70	4	17	0.115422

As this test scenario produced the optimal results; thus, the clustering algorithm predicted class labels were mapped to the actual label of the dataset. Mapping cluster labels to original data labels is a crucial step in the result analysis. It allows the clustering accuracy to be expressed in terms of accuracy used in supervised algorithms.

To visualize the data, we generated a scatter plot and confusion matrices. The scatter plot in figure 21 and figure 22 shows the actual domain labels and predicted domain labels against the content numbers for K-Means and DBSCAN. Table 15 and Table 16 show the confusion matrix results of the clustering algorithm after the mapping. The confusion matrix shows the predicted labels against the actual labels of the dataset. The domains entertainment, environment, live living, sports, and tech news show promising results as the algorithms correctly predicted most class labels. Interestingly, the prediction level in the youth and business domains performs imperfectly in both the K-means and DBSCAN algorithms, although these two algorithms have different properties. In the research study of [19], the feasibility of a document representing various domains was assessed. the feasibility of a document representing various domains was assessed. We have used a domain expert's help to evaluate the categorization results in this study, which resulted in clear speculation that a single document can identify in multiple domain names. This study can determine why two domains, i.e., "youth" and "Business," are predicted to be more spread out in the clusters.

4.3 Clustering on Daily Star Dataset

DailyStar Mapping Contents with actual Domain and Clustered labels(Predicted Domain) KMeans

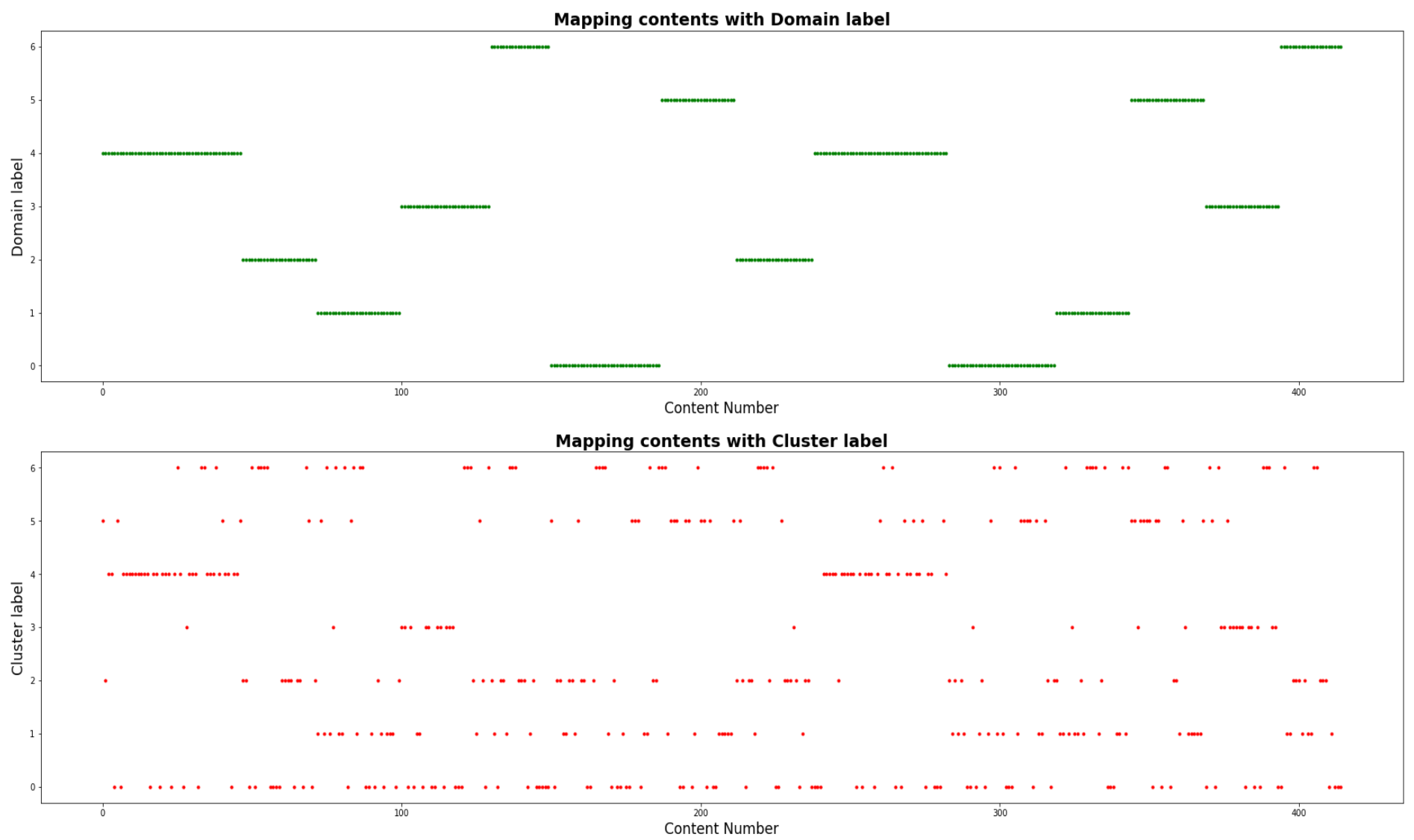


Fig. 21. Mapping Clusters with Domain of KMeans Result

4.3 Clustering on Daily Star Dataset

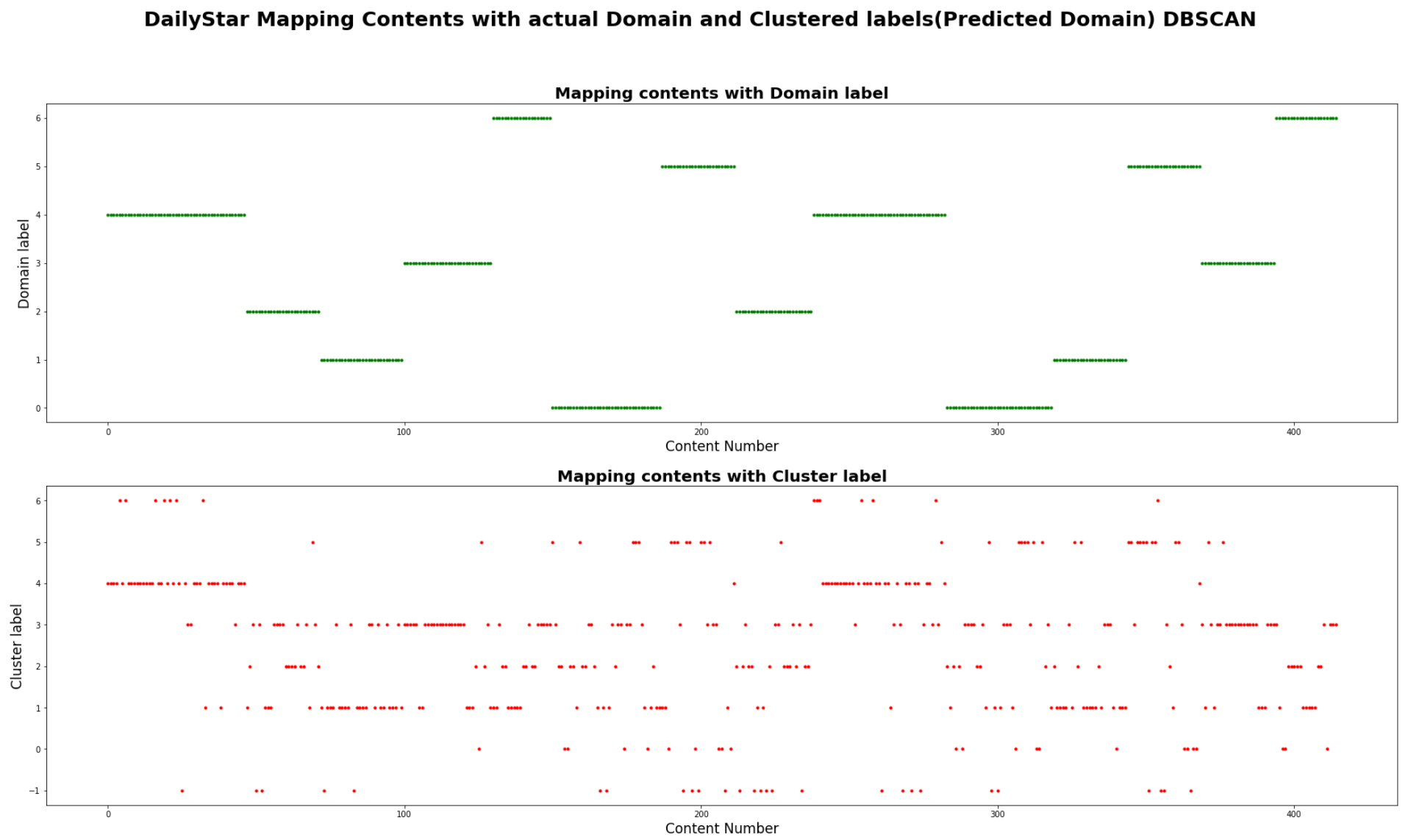


Figure 22. Mapping Clusters with Domain of DBSCAN Result

Table 15: Confusion Matrix of KMeans on Daily Star Dataset

Predicted Domain-Cluster Mapped Labels	Actual Domain Labels						
	business	entertainment	environment	life living	sports	tech news	youth
business	18	9	14	16	20	9	12
entertainment	17	21	2	3	0	13	9
environment	16	5	20	2	2	2	14
life living	1	2	1	22	1	2	0
sports	0	0	0	0	54	0	0
tech news	12	2	3	3	9	19	0
youth	9	14	11	9	6	5	6

Table 16: Confusion Matrix of DBSCAN on Daily Star Dataset

Predicted Domain-Cluster Mapped Labels	Actual Domain Labels						
	business	entertainment	environment	life living	sports	tech news	youth
business	9	1	0	1	0	9	3
entertainment	14	34	7	11	3	4	13
environment	15	3	19	2	0	1	13
life living	19	11	15	38	9	7	12
sports	0	0	0	0	61	2	0
tech news	12	2	2	3	1	18	0
youth	0	0	0	0	13	1	0
Non-clustered	4	2	8	0	5	8	0

The performance analysis of K-Medoid has not been included due to the inferior quality of the clusters. The K-Medoid algorithm is not suitable for large text document clustering because of its many dimensionalities. The actual document might not represent the cluster correctly. In K-Means, the average score is calculated to produce clusters. Therefore, a document representing multiple domains gets a higher weight, and a cluster tends to be more accurate while being formed. However, K-Medoid, which uses an actual random point as a medoid, frequently fails to cluster documents representing different domains.

4.3.1 Comparison Between KMeans and DBSCAN result on Daily Star:

Table 17: Comparison between KMeans and DBSCAN clustering Result

	Precision		Recall		Accuracy	
	KMeans	DBSCAN	KMeans	DBSCAN	KMeans	DBSCAN
business	0.184	0.391	0.247	0.123	0.386	0.431
entertainment	0.323	0.395	0.396	0.642		
environment	0.328	0.358	0.392	0.373		
life living	0.759	0.342	0.400	0.690		
sports	1.000	0.968	0.587	0.663		
tech news	0.396	0.474	0.380	0.360		
youth	0.100	0.000	0.146	0.000		

4.3 Clustering on Daily Star Dataset

Table 17 shows the precision-recall and accuracy comparison of the KMeans and DBSCAN algorithms, based on tables 15 and 16 which are the confusion matrix for these two algorithms. For this study, we have to give more importance to the Recall score measurement. The recall is a metric that quantifies the ratio between correct positive among correct positive and false negative predictions made. As one single document can represent multiple domains, the number of false positives will not have a negative effect on the accuracy. High precision relates to the low false positive rate. Recall provides an indication of missed positive predictions. When a user wants to search for similar documents, missed positive documents indicate a bad system. When comparing the accuracy scores of KMeans and DBSCAN, it is clear that DBSCAN algorithms are more exact. The domain 'entertainment', 'sports', and 'Life_living' gave out fewer false negative predictions in both of the algorithms.

4.4 Key Findings

The task of clustering a set of big textual documents so that texts in the same group are more contextually similar is known as text document clustering. Clustering huge text documents have been a major area of research in NLP, as contextual word embeddings are required for clustering.

Traditional text document classification methods represent documents with non-contextualized word embeddings and vector space models. The existing text document classification methodologies have been explored first and then we evaluated their strengths and limitations.

The study started with models based on Bag-of-Words and shifted towards transformer-based architectures. Several classification algorithms have been applied to the word embeddings of the pre-trained state-of-art BERT model.

The findings from this capstone project:

- It is concluded that transformer-based embedding is necessary to capture the contextual meaning.
- To build this framework all the latest technologies and state-of-art tools have been used.
- The state of art BERT model has been used in this capstone project.
- BERT is mostly used for next word prediction but here it's used for a different approach which is text classification and clustering.
- For classification, by comparing this work with Yutong Li et al. [17], it has been found that this study and experiment show better accuracy than target work. The study results have outperformed the results of Yutong Li et al. [17]. with 50% improvement.
- Among all three state-of-art classification algorithms – Decision Tree, Random Forest, and K-Nearest Neighbor, random Forest performed better than others for every dataset as it is an ensemble method.
- This research has applied two clustering algorithms- KMeans and DBSCAN, over the Daily Star dataset to measure the performance of the algorithm over BERT's generated contextual embeddings. After the clustering of the embedding produced by BERT, as clustering is an unsupervised algorithm and obtaining the accuracy is often difficult, a mapping method to evaluate the clustering performance is implemented. A new algorithm has been implemented for cluster domain mapping.
- For better visualization of the data, a scatter plot and confusion matrices have been generated. This different approach of plotting is used for mapping contents with the actual domain and clustered labels.
- From table 16, it is visible that prediction levels are more correct when applying DBSCAN. This is due to DBSCAN being a density-based clustering method and taking account of the document vector's density rather than mean values are the more suitable approach for document clustering.

Conclusion & Future Work

Our capstone project introduces a novel text classification framework and text clustering method with a mapping algorithm using transformed-based embeddings. The study of embedding schemes of different models concluded that the transformer-based architectures are better suited for generating contextual word embeddings.

This project has applied the classification algorithms to a benchmark Reuters dataset and web-scraped Daily Star dataset. The experiments with the random forest classifier and KNN show promising results. The random forest classifier achieves around 90% accuracy over the Reuters dataset and 75% accuracy over the Daily Star dataset. The performance of the classifiers has been compared with existing work and it shows up to 50% improvement in accuracy. Web-scraped Daily Star dataset was used in this project to test clustering techniques. According to the K-Means clustering experiment, cluster number 5 has the highest Dunn index and Silhouette Coefficient, indicating this cluster is more compact and well segregated. Furthermore, the experiment with the confusion matrix of the DBSCAN and K-Means algorithm shows which classes accurately predicted most class labels. Entertainment, environment, live living, sports, and tech news show promising results. The mapping algorithm has made it possible to evaluate the unsupervised clustering algorithm in a supervised way, validating the accuracy more precisely and satisfactorily. The comprehensive performance study reveals that K-Means and DBSCAN can cluster most of the dataset correctly.

We have extended our study towards building a user-preferred document retrieval system. In the sign-up page all the list of domains and topics are given, and users can select them. After this selection their preference will be stored and based on this history the clustering result will be generated. With this, the domains become more concise. The extended version of the framework will include a hybrid recommender system integrating user-centric preferences. The hybrid recommender system facilitates both content-based and collaborative filtering. Moreover, multiple repositories of text documents can be added to the framework for a smooth user experience of information browsing. The framework will eventually allow users to browse and submit documents for finding contextually similar records. The framework will be more robust as the time passes by through a monitoring scheme, keeping the updated user profiles based on the search history.

References

- [1] Y. Li, J. Cai and J. Wang, "A text document clustering method based on weighted bert," 2020..
- [2] S. Yeasmin, N. Afrin, K. Saif and M. R. Huq, "Development of a Text Classification Framework using," in *Data2022*, Lisbon, Portugal, 2022.
- [3] A. Fawcett, "Data science in 5 minutes:What is one hot encoding?," [Online]. Available: <http://www.educative.io/blog/one-hot-encoding>.
- [4] O'Reilly, "Text Vectorization and Transformation Pipelines," [Online]. Available: <https://www.oreilly.com/library/view/applied-text-analysis/9781491963036/ch04.html>.
- [5] "Nlp concepts:Word-embeddings," [Online]. Available: [//medium.com/@data_datum/nlp-concepts-word-embeddings-99337bf98b3](https://medium.com/@data_datum/nlp-concepts-word-embeddings-99337bf98b3),.
- [6] "Word2Vec," [Online]. Available: <https://en.wikipedia.org/wiki/Word2vec>..
- [7] "Word embeddings and their challenges," [Online]. Available: <http://aylien.com/blog/word-embeddings-and-their-challenges>..
- [8] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
- [9] Y. Kim, C. Denton, L. Hoang and A. M. Rush, "Structured attention networks," *arXiv preprint arXiv:1702.00887*, 2017.
- [10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, K. N, A. Gomez and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, 2017.
- [11] G. Giacaglia, "Towards Data Science," 11 March 2019. [Online]. Available: <https://towardsdatascience.com/transformers-141e32e69591>.
- [12] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2019.
- [13] G. Ghati, " Comparison between bert, gpt-2 and elmo," [Online]. Available: <http://medium.com/@gauravghati/>.
- [14] Y. Liu, M. Ott, G. Myle , N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Veselin, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [15] f. Team, "fairseq Team by Facebook AI. Roberta—pytorch," [Online]. Available: http://pytorch.org/hub/pytorch_fairseq_roberta.
- [16] V. Sanh, L. Debut, J. Chaumond and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.
- [17] Y. Li, J. Cai and J. Wang, "A text document clustering method based on weighted Bert mode," *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, 2020.
- [18] W. Hu, D. Xu and Z. Niu, "Improved K-Means Text Clustering Algorithm Based on BERT and Density Peak," *2021 2nd Information Communication Technologies Conference (ICTC)*, 2021.
- [19] H. Shi and C. Wang, "Self-supervised document clustering based on bert with data augment," *arXiv preprint arXiv:2011.08523*, 2020.

- [20] W.-C. Chang, H.-F. Yu, . K. Zhong, Y. Yang and I. Dhillon, "X-bert: extreme multi-label text classification with using bidirectional encoder representations from transformers," *arXiv preprint arXiv:1905.02331*, 2019.
- [21] A. Ashutosh , A. Ram, R. Tang and J. Lin, "Docbert: Bert for document classification," *arXiv preprint arXiv:1904.08398*, 2019.
- [22] D. Godoy and A. Amandi, "PersonalSearcher: An Intelligent Agent for Searching Web Pages," *Advances in Artificial Intelligence, International Joint Conference, 7th Ibero-American Conference on AI, 15th Brazilian Symposium on AI, IBERAMIA-SBIA 2000, Atibaia, SP, Brazil, November 19-22, 2000, Proceedings*.
- [23] A. Subakti, H. Murfi and N. Hariadi, The performance of BERT as data representation of text clustering, SpringerOpen, 2022.
- [24] R. L. a. C. Group, "Reuters," [Online]. Available: <https://www.kaggle.com/nltkdata/reuters>.
- [25] S. Yeasmin, N. Afrin, K. Saif and M. R. Huq, "Daily star dataset.," [Online]. Available: <https://github.com/Sumona062/Daily-Star-Datasets>.
- [26] [Online]. Available: <https://www.projectmanager.com/guides/work-breakdown-structure>.
- [27] [Online]. Available: <https://uwaterloo.ca/ist-project-management-office/methodologies/project-management/planning/work-breakdown-structure/wbs-benefits>.
- [28] [Online]. Available: <https://dayshape.com/what-is-resource-allocation-benefits-importance-methods>.
- [29] P. Kukhnavets, "Hygger," [Online]. Available: <https://hygger.io/blog/what-is-critical-path-method-for-in-project-management/>.

Appendix

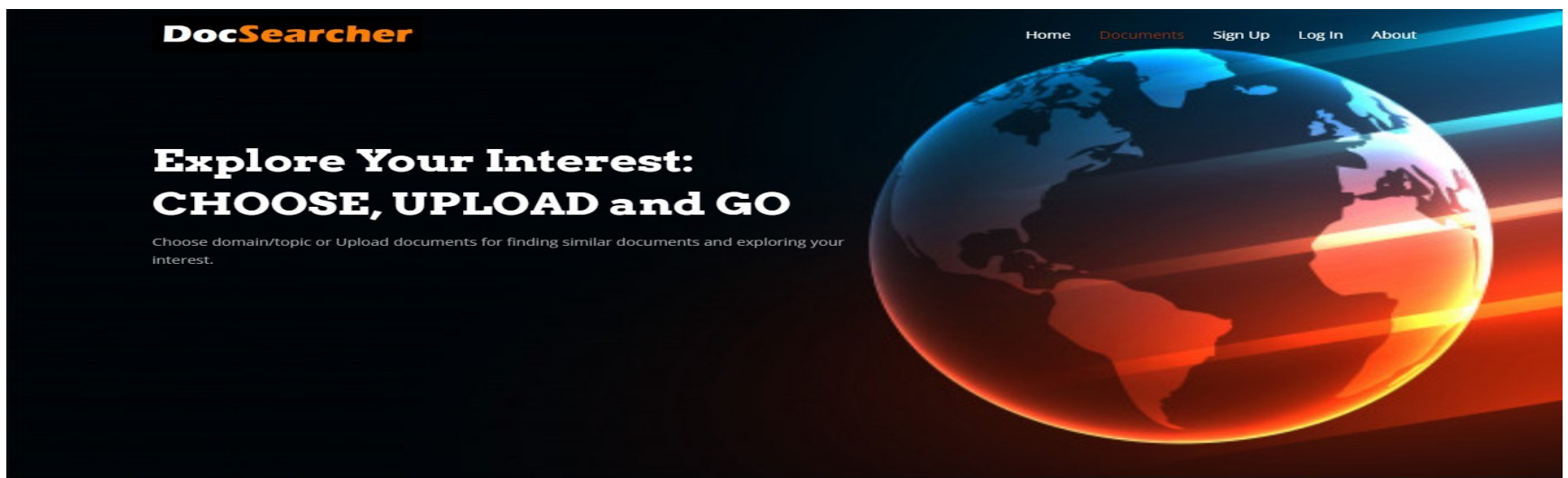
Appendix A

Addressing COs, Knowledge Profile (K), Complex Engineering Problem (EP), and, Program Outcome(PO):

CO	CO Descriptions	K	EP/EA	PO
CO7	Assess and address the sustainability and impact of the capstone project in societal and environmental contexts	K7: Our solution provides safety to the user's data and doesn't store any sensitive information which can create unrest between the religious or political groups. Also, to ensure sustainability we used google colab as it provides cloud-based service.	EP5: For our capstone project, we have used the concept of ethical Machine Learning. We have followed the data life cycle's ethical Machine Learning's standard processes. EP6: We need to train our model based on the user perspective and domain level. But we have to filter sensitive content so that it can create fuss between different social groups.	PO7: We have maintained maximum sustainability by implementing cloud-based paperless services. We are sharing services over a network to maximize the effectiveness of the resources thus maximizing the computing power and reducing environmental damage. We are using google colab as our implementing platform. Google appreciates and encourages sustainability and cloud computing for a sustainable environment and technologies. The main motivation behind the clustering and classification framework is that it's automated so it's a paperless service.
CO8	Apply professional and engineering ethical principles and practices for the implementation of the capstone project	K7: The developed solution provides security of the user's data and preferences to feel safe using our application. Furthermore, we have used colab from google. Google appreciates and encourages sustainability and cloud computing for a sustainable environment and technologies.		PO8: In the corpus of text clustering framework, we are not storing sensitive information. The prototype does not introduce any political or religious bias. We have filtered the sensitive document
CO9	Work effectively as an individual and a team member for the successful completion of the capstone project			PO9: Throughout this capstone project, we have had effective meetings with the supervisor as well as individually. We've divided our work, and, in the end, we've sat together to combine everything so that we

				can achieve the main objectives of the capstone project. The communication and understanding between the three of us have become stronger now than at the beginning of the project.
CO10	Write effective reports and design documentation, and make effective presentations of the outcome of the capstone project		<p>EA1: As our resources, we have studied the relevant domain.</p> <p>EA2: In this project, there are different stakeholders with different expertise and intellectual level. Therefore, we built the framework in such a way that can provide understandable results for different stakeholders.</p> <p>EA3: A unique, innovative, robust, sustainable, environment-friendly, user-friendly platform consisting of all the state of art technologies.</p> <p>EA4: We have trained our model based on the user perspective and domain level. Some users may require data from a quarsal level of the domain. More specialized users may want to find documents at a finer level of granularity.</p> <p>EA5: Human interpretation of the documents can be varied and it's very important for us to tag the documents. That is an unfamiliar issue for us because we lack expertise in linguistics. That's why we have consulted with a domain expert.</p>	PO10: For our capstone project, we have prepared an informative slide for better understanding. We have also prepared the report in an organized way and followed a proper flow so that it can be easier to understand. Lastly, we have tried to write the report with minimum grammatical errors. For that, we have checked it using Grammarly.

CO11	<p>Conduct economic analysis and cost estimation; and apply appropriate project management processes in the development life cycle of the capstone project</p>			<p>After the economic analysis, our projected total budget is around BDT 7.5 lakhs. We have made a work breakdown structure for identifying different task-wise budgets. We've made resource planning of the project to get a better estimation of the timeline of the project as well as the dependencies of each task. We have identified the critical paths for the project using the critical path method(CPM). We have also calculated the break-even point. After around 7 months later we expect to get the revenue.</p>
CO12	<p>Prepare to take part in independent and lifelong learning for adapting emerging technologies for the solution of the complex computer science and engineering problems</p>			<p>For this project, we have studied many existing technologies and did a literature survey, also experienced of conducting research that provides a viable product that helps our lifelong learning.</p> <p>After this research, now we know different steps of conducting research and how research can be translated into a product. So we are confident that in the future we would always have an inquisitive mind to tackle any problem and come up with a viable solution that can be also transformed into a hardware or software product.</p>

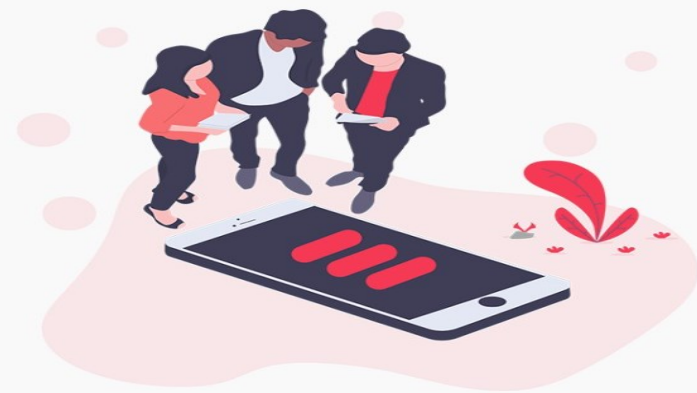


ALL DOCUMENTS


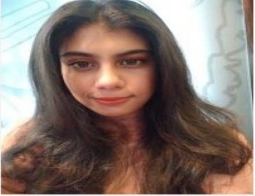



About This Website

Docsearcher is a multi-domain similar text retrieval framework, offering users to find similar documents conveniently. Docsearcher is designed to meet the information need of every professional. Users can choose a domain and dataset or browse through documents without hassle. Docsearcher has multiple repositories users can choose from to retrieve the equivalent document. Records are retrieved with the Machine Learning backend. Docsearcher also recommended similar documents to what the user uploads. Enjoy the unique features with only a click. Upload, select, and go.



Meet Our Team

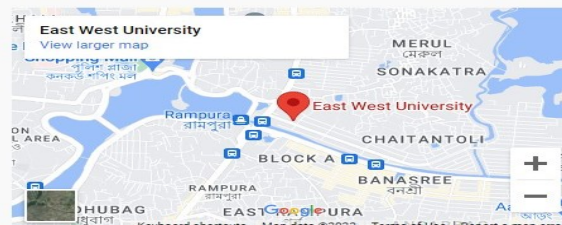
	<p>SUMONA YEASMIN Student Education: B.Sc in Computer Science and Engineering, East West University, Dhaka, Bangladesh Email: 2018-2-60-062@std.ewubd.edu, sumonasumu930@gmail.com</p> <p>f t</p>		<p>NAZIA AFRIN Student Education: B.Sc in Computer Science and Engineering, East West University, Dhaka, Bangladesh Email: 2018-2-60-023@std.ewubd.edu, nowshinwork3247@gmail.com</p> <p>f t</p>
	<p>KASHFIA SAIF Student Education: B.Sc in Computer Science and Engineering, East West University, Dhaka, Bangladesh Email: 2018-2-60-001@std.ewubd.edu, ksalf2010@gmail.com</p> <p>f t</p>		

Contact Us

Email:
sumonasumu930@gmail.com
nowshinwork3247@gmail.com
ksalf2010@gmail.com

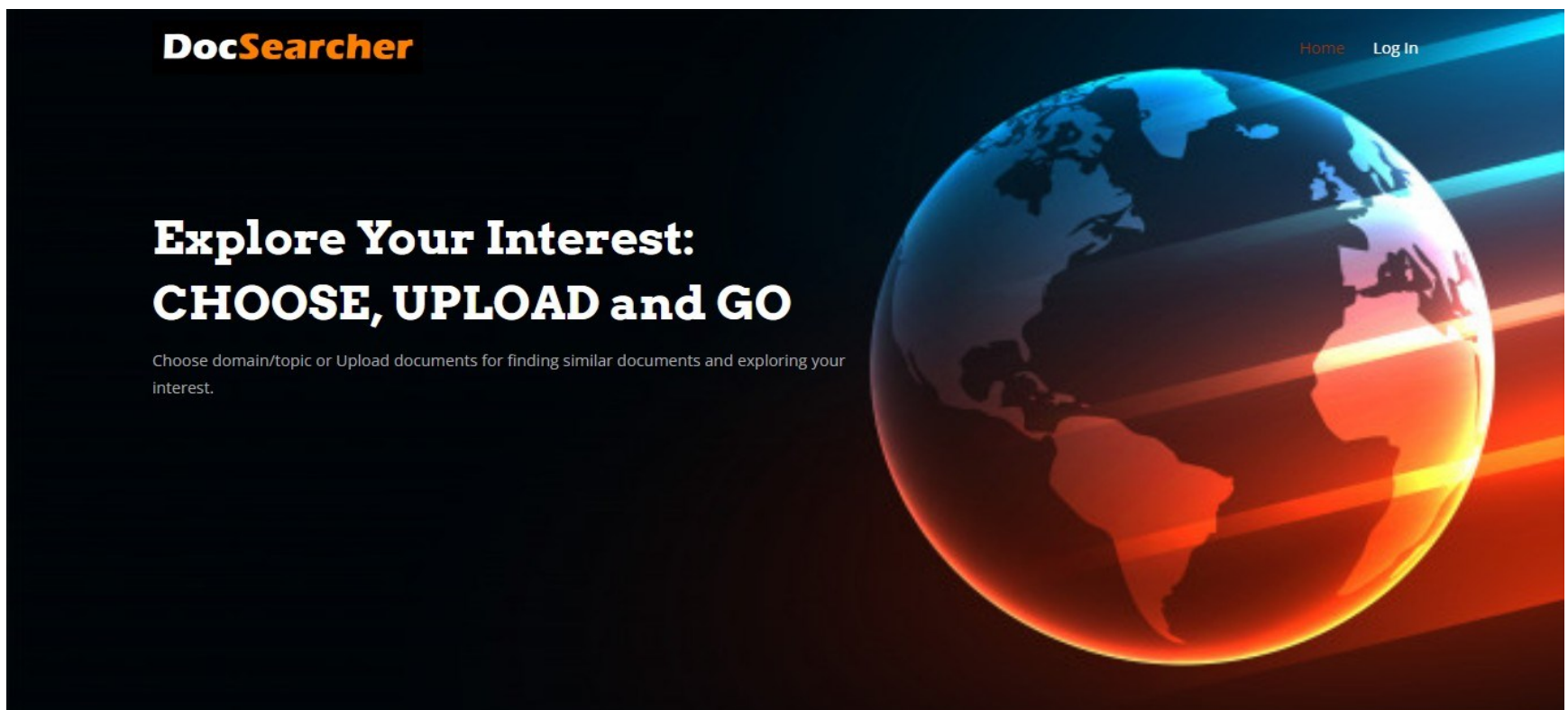
Address: East West University, Plot No-A/2, Jahurul Islam City, Aftabnagar, Dhaka-1212, Bangladesh

Contact No:
01992616930
01521515419
01977448080



Home

Home page of the DocSearcher Prototype. This page is accessible for registered and non-registered users. Users can view the existing repositories, know about the developer team, contact the developers from this page, and learn what this application is about.



Create Your Profile

<input type="text" value="Name"/>	<input type="text" value="Email"/>
<input type="password" value="Password"/>	<input type="password" value="Re-Type Password"/>
<p>Select Interested Domains for DailyStar Repository</p> <div style="border: 1px solid #ccc; padding: 2px;"><ul style="list-style-type: none">BusinessEnvironmentEntertainmentSportsLife Living</div>	<p>Select Interested Topic for Reuter Repository</p> <div style="border: 1px solid #ccc; padding: 2px;"><ul style="list-style-type: none">CocoaEarnings and Earnings ForecastsMergers/AcquisitionsCopperCoffee</div>
<input type="submit" value="SUBMIT"/>	

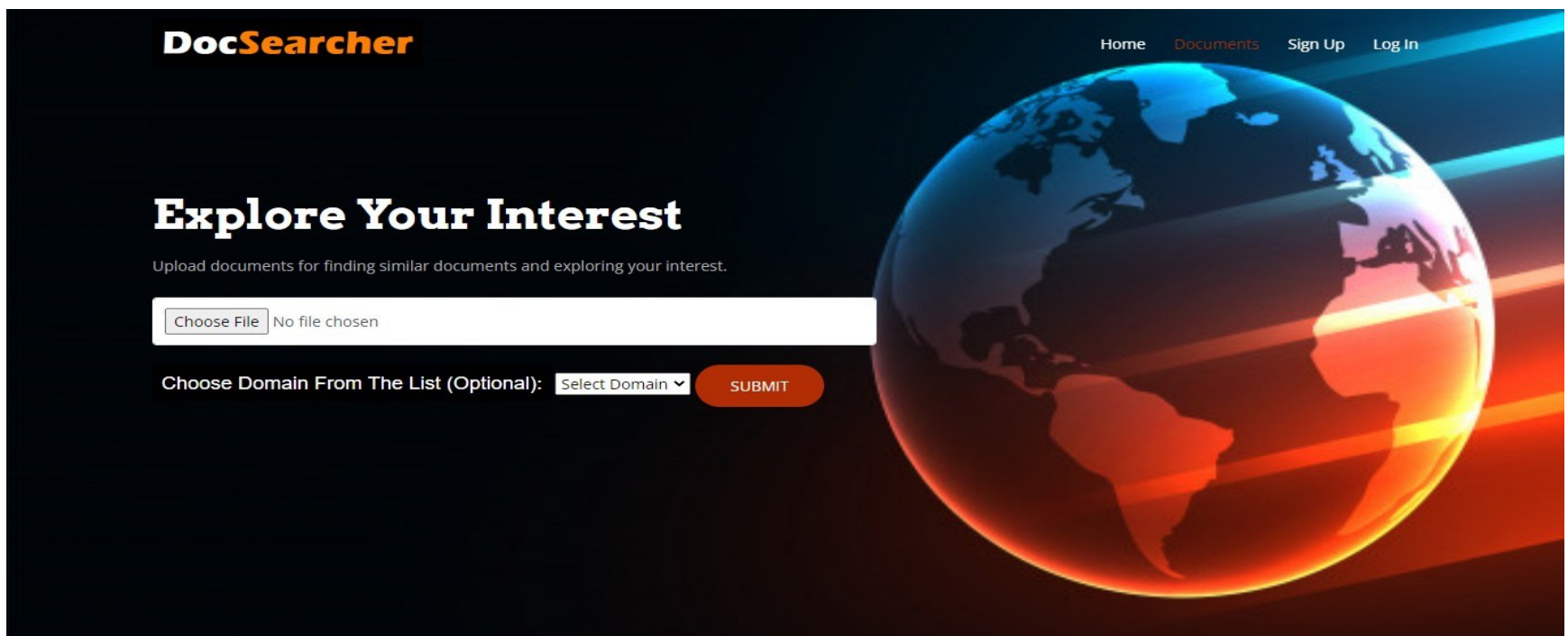
Sign Up

Protypes's sign-up page. Users can sign up to enjoy user-centric document searching via this page. Users can choose multiple domains or topics from the existing repositories as their interest areas. This data is utilized to produce a user-centric search system for that particular user. This is entirely optional. Users may not choose an attractive size if they don't want it.

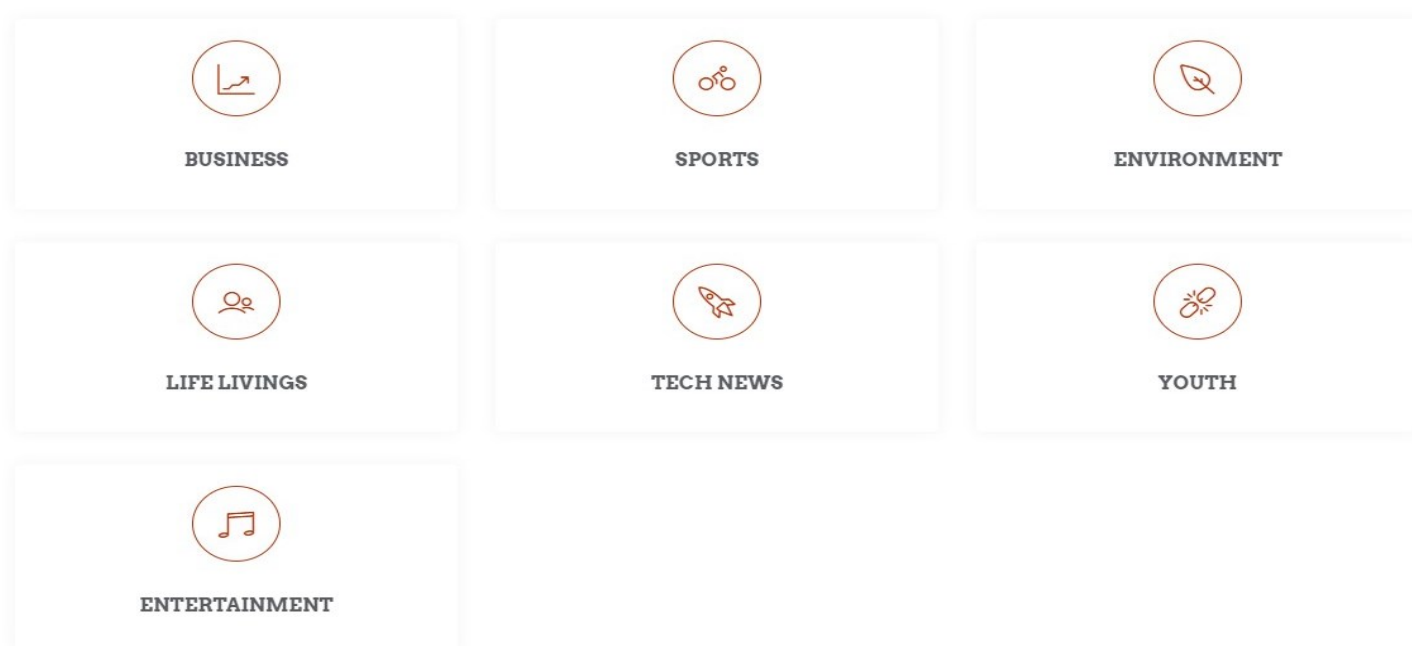


Login

Login system of the DOcSearcher, Already registered users can log in at any time to use the system and browse through repositories; upload a document, and find similar ones based on the areas of their interest.



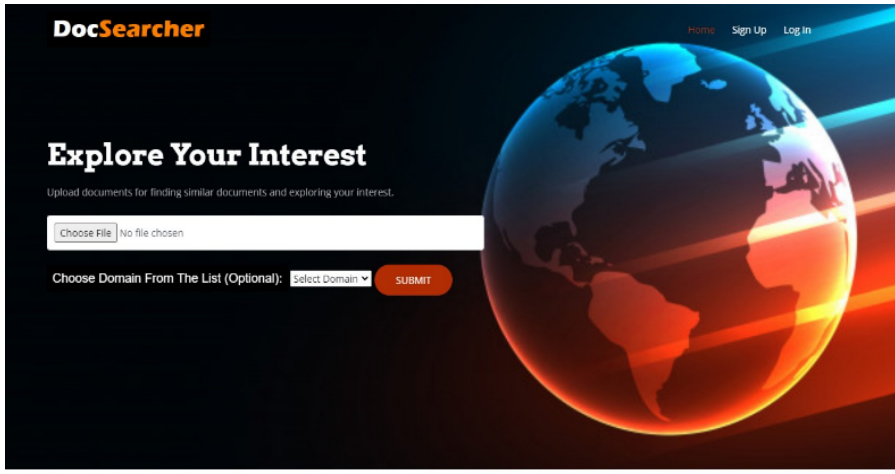
DAILY STAR DOCUMENTS



Daily Star

In the Daily Star repository, registered and non-registered users, can view this page and browse through the seven domains and their contents. Users can also upload a document to find similar ones. In that case, user can enjoy two options: Domain dependant and Independent searching. Users with more knowledge of the domains can select a domain for searching, whereas new users can also search without selecting any domain from The repository Daily Star.

Appendix B



Document ID: 0

Rafael Nadal said Monday he has tested positive for Covid-19 on his return to Spain from Abu Dhabi where he took part in an exhibition tournament last week. "I am now confined at home and have informed the people who had been in contact with me," he wrote on Twitter. The Spanish former world number one had been due to fly to Melbourne later this month to compete in an ATP event ahead of the Australian Open.

Document ID: 1

14-day self-isolation of two cricketers of the Bangladesh women's team came to an end as Bangladesh Cricket Board's (BCB) chief physician Dr. Debashish Chowdhury confirmed that their recent Covid-19 tests returned negative today. These cricketers had tested positive for coronavirus after returning from Zimbabwe earlier this month following the World Cup qualifiers, which was halted midway due to emergence of a new Covid-19 variant in Africa. "Both cricketers returned negative and they are okay. They will go back to their respective homes," Dr. Debashish stated. These two cricketers were the first victims of Omicron (the latest variant of coronavirus) in Bangladesh. Fortunately, they had mild symptoms and did not face any physical difficulties. Now the clearance enables these Tigrisses to socialise after a couple of weeks of quarantine. Earlier, the rest of the Tigrisses were moved to the BCB Academy where they had undergone 14-day mandatory quarantine since returning to country. Followed by their latest Covid-19 test results [which all returned negative], they were discharged and sent home ahead of Victory Day last week.

Document ID: 2

Australia beat England by 275 runs in the day-night second Ashes Test at the Adelaide Oval on Monday to go 2-0 up in the five-match series. Chasing a daunting victory target of 468, England were all out for 192 in the final session of the pink-ball contest. Chris Woakes' 44 was the highest individual score in England's abject second innings batting display, while Jos Buttler frustrated Australia for a while with a dogged 26 off 207 balls. Jhye Richardson was the pick of the Australian bowlers, claiming 5-42. The third Test begins in Melbourne on Sunday.

Document ID: 3

As build-up to the upcoming ICC Under-19 Cricket World Cup in the West Indies ends next year, the Bangladesh Under-19 team are scheduled to depart for UAE tomorrow to take part in the ACC Asia Cup. Bangladesh will feature in group B of the tournament and kick off campaign against Nepal on December 24 at the Sharjah Cricket Stadium. Rakibul Hasan and his troop will also face Kuwait and Sri Lanka on December 25 and 28, respectively, in the group stage. A lot will be expected from Bangladesh, even though the defending champions could not engage in the best of preparations ahead of the World Cup compared to the previous edition. Prior to their successful World Cup campaign last year, the young Tigers played 35 matches, whereas the current team have played just 13 games before heading to the Asia Cup this time around. Head coach Navid Newaz is not too bothered about the comparative lack of preparation and the pressure of defending the title. "Not exactly [feeling pressure] because we have been with this group for almost one year now with pandemic scenario, in and out. Besides, we have come to an understanding that this is the way we are going to play our cricket," Navid told the media today. "And being the defending champions last year, we are prepared to use the pressure [of title-defense] to our advantage instead of feeling it [negatively]. We have to make sure we take every game one at a time, which is the most important thing going into a tournament," the head coach stated. The U-19s, after completing the Asia Cup, will travel directly to the West Indies from the UAE to take part in the ICC Under-19 World Cup, scheduled to begin from January 15.

Document ID: 4

There have been insinuations, remarks and opinions regarding all-rounder Shakib Al Hasan's involvement in Test cricket from the media, cricket experts and even board officials. Looking back, Shakib had missed a few Tests in recent years, lending credibility to arguments about his lack of interest in Tests. Opting out of the New Zealand tour and the consequent drama has put these arguments back in focus. However, there exist firmer points between the two arguments that 'Shakib doesn't want to play 'Tests' and 'Shakib doesn't enjoy Tests'. The second Test against Pakistan in Dhaka, which Bangladesh came very close to grinding out a draw from yesterday, showcased Shakib's enjoyment but also the demotivating factors. Shakib played one of his very special knocks under pressure, but in the end the result did not go in Bangladesh's favour. When that happens, how does Shakib feel? Shakib's ability to not allow anyone to dictate terms was once again on display as he did not let the bowlers feel they were on top of him for a moment. How he dealt with spinners was exemplary. The Pakistan spin duo had a general plan for the Bangladesh batters, getting the ball to pitch right in front of the bat and getting that slight turn to do the trick for them. Shakib was able to adjust remarkably within the same delivery, from a front foot to a slightly back foot stance. He did not need to release shots like Mushfiqur Rahim for instance, but was able to make adjustments late to keep scoring. From a pressure situation, Shakib had a groove about him which took Bangladesh close to a result. A pivotal moment came when Mehedi Hasan Miraz got out to another sweep in the second innings which saw Pakistan come back into the game. As soon as the umpire had signaled not out and Pakistan went for the review, Shakib vehemently chastised Mehedi for playing that shot. Mehedi's two dismissals were significant to the fate of the match. Without arguing in Shakib's favour, when a good player sees the supporting cast fall alongside the team processes, any player has a tendency to be demotivated. Shakib's long-term mentor Nazmul Abedeen Fahim analysed the scenario for 'The Daily Star'. "In the first innings, he counter-attacked to avoid the follow-on due to match scenario. In the second innings, I was relieved to see that he wanted to play good cricket. What he did was very difficult because you need to be in the groove for a long period of time. Unless you love the game and the challenge, you can't do that. One of his best innings," Fahim said. "If he was careless, he wouldn't be able to play like that," he said. However, once Miraz got out, Shakib changed too, even playing a lofted shot over mid-off that fell short, pointing to his dwindling morale. "Take the case of Mahmudullah Riyad. He was felt some 17 months before his Test retirement that he wouldn't be considered [for Tests]. How demotivated must he have been? But his decision was like shooting yourself in the foot. It became an emotional thing and players are going through such demotivation at the moment," Fahim dictated. It cannot be argued that there exists a gap in quality between the generation of the Big Five and the current crop of players. When there are no correct procedures and the existence of haphazard selection process alongside bad results on the pitch, players will find a way to relieve workload or responsibility. That may very well be the phase Shakib's Test career is going through. Not a lack of enjoyment but an existence of demotivation.

Document ID: 5

Tottenham Hotspur's European campaign for this season is over after UEFA awarded Rennes a 3-0 victory in their final Europa Conference League group match on Monday. Spurs were unable to fulfil the December 9 fixture due to a swathe of Covid-19 cases in the squad. As a result they finish third in the group behind group winners Rennes who progress directly to the last 16 and Dutch side Vitesse who now face Rapid Vienna in a knockout play-off. The Uefa Appeals Body took the following decision: to declare the 2021/22 UEFA Europa Conference League group stage match between Tottenham Hotspur FC and Stade Rennais FC, that was initially scheduled to be played on 9 December 2021, as forfeited by Tottenham Hotspur FC, who is therefore deemed to have lost the match 0-3. "read the UEFA statement. The match in London had to be called off when Spurs manager Antonio Conte announced on the eve of the game eight of his players and five members of staff had tested positive for Covid-19. Spurs have been one of the hardest hit English clubs by the resurgence of Covid-19 – largely due to the Omicron variant – having had to postpone league games against Brighton and Leicester as well. However, unlike many of their domestic rivals, they were able to fulfil their fixture at the weekend – a 2-2 draw with Liverpool.

Document ID: 6

"The pic you've been waiting for," read the post from Barcelona's official Twitter account last month, above a photo of the team's new coach Xavi Hernandez talking to a smiling Frenkie de Jong. Xavi was barely into his second week as coach but the hope was that he would already have De Jong nestled under his wing, the advice flowing from one midfielder to another and the clock counting down until the flourishing of another Dutch master in Barcelona colours. For three years now, including two in Spain and his final one at Ajax, the destiny of De Jong has, through no fault of his own, defined almost everything about him as a footballer. His performances have been up and down, and mostly somewhere in between. He has been an obvious pick in Barca's starting line-up but only occasionally a stand-out performer. Against Elche at Camp Nou on Saturday, he was substituted in the second half and whistled by some of his own fans. "Maybe you don't see what he does but he's very good," said Xavi. "I don't think I have ever done really badly, but I know I can do better," De Jong told AFP in an interview last year. In some ways, the promise of what De Jong can be has been all-consuming. Honed by Ajax and instrumental for the Dutch national team, De Jong owned two fundamental characteristics to be considered a carrier of 'Barca DNA': As a child, he even did multiple stadium tours of Camp Nou and wore a shirt with Lionel Messi's name on the back. Perhaps the only way the anticipation could grow further was through a transfer scramble, which was provided when Paris Saint-Germain and Manchester City joined the race for De Jong's signature in 2018. It was a fight Barcelona could not afford to lose. Excitement around De Jong had been building ever since he helped engineer Ajax's dismantling of Real Madrid in the Champions League, while Barca's own failures in the tournament meant the craving for young talent had morphed into something closer to desperation. There was political pressure too. The club's president Josep Maria Bartomeu faced anger about repeated capitulations and an increasingly unrecognisable style of play, all on the back of Neymar's humiliating departure to PSG - poster boy - The prospect of PSG, a state-funded club, with its perceived lack of identity or history, again beating Barcelona to a player like De Jong, a poster boy for almost everything Barca stood for, was unthinkable. Barcelona paid 75 million euros for De Jong, with a further 11 million euros due in variables, and a club announcement explained in its third paragraph how "the president Josep Maria Bartomeu and CEO Oscar Grau were personally involved in closing the deal". "In some ways I've always been a Barcelona player," said De Jong, with Camp Nou behind him. Yet in the rush to be seen to buy young, bolster the team's identity, beat PSG and boost the popularity of the president - what was Barcelona's plan for De Jong? How did they expect to improve or even accommodate this talented 21-year-old? In the first game of the season, Barcelona's then-coach Ernesto Valverde offered a show of faith by fielding him ahead of Sergio Busquets. But Barca struggled, Busquets was restored and De Jong shifted to the right of the midfield three. In his two and a half years, De Jong has played right, left, defensive and attacking midfield, as well as right-wing and centre-back. He has played under four permanent coaches, seen Messi leave and Barcelona surrender its status as a serious European club. His best position remains unclear and his future is now in question. Barcelona's financial meltdown means any significant sale is always tempting and De Jong has been linked to both Manchester clubs in recent weeks. He would likely fetch more than 50 million euros in either January or the summer transfer window. For now, the onus is on De Jong to take what could be his last chance under Xavi, to satisfy lofty expectations and make

Documents for “Sports” Domain

Document Present inside the Daily Star repositories Sport’s Domain. Inside each domain, the

existing contents can be viewed with the document id and the text.

Document ID: 6

"The pic you've been waiting for," read the post from Barcelona's official Twitter account last month, above a photo of the team's new coach Xavi Hernandez talking to a smiling Frenkie de Jong. Xavi was barely into his second week as coach but the hope was that he would already have De Jong nestled under his wing, the advice flowing from one midfielder to another and the clock counting down until the flourishing of another Dutch master in Barcelona colours. For three years now, including two in Spain and his final one at Ajax, the destiny of De Jong has, through no fault of his own, defined almost everything about him as a footballer. His performances have been up and down, and mostly somewhere in between. He has been an obvious pick in Barca's starting line-up but only occasionally a stand-out performer. Against Elche at Camp Nou on Saturday, he was substituted in the second half and whistled by some of his own fans. "Maybe you don't see what he does but he's very good," said Xavi. "I don't think I have ever done really badly, but I know I can do better," De Jong told AFP in an interview last year. In some ways, the promise of what De Jong can be has been all-consuming. Honed by Ajax and instrumental for the Dutch national team, De Jong owned two fundamental characteristics to be considered a carrier of 'Barca DNA': As a child, he even did multiple stadium tours of Camp Nou and wore a shirt with Lionel Messi's name on the back. Perhaps the only way the anticipation could grow further was through a transfer scramble, which was provided when Paris Saint-Germain and Manchester City joined the race for De Jong's signature in 2018. It was a fight Barcelona could not afford to lose. Excitement around De Jong had been building ever since he helped engineer Ajax's dismantling of Real Madrid in the Champions League, while Barca's own failures in the tournament meant the craving for young talent had morphed into something closer to desperation. There was political pressure too. The club's president Josep Maria Bartomeu faced anger about repeated capitulations and an increasingly unrecognisable style of play, all on the back of Neymar's humiliating departure to PSG - poster boy - The prospect of PSG, a state-funded club, with its perceived lack of identity or history, again beating Barcelona to a player like De Jong, a poster boy for almost everything Barca stood for, was unthinkable. Barcelona paid 75 million euros for De Jong, with a further 11 million euros due in variables, and a club announcement explained in its third paragraph how "the president Josep Maria Bartomeu and CEO Oscar Grau were personally involved in closing the deal". "In some ways I've always been a Barcelona player," said De Jong, with Camp Nou behind him. Yet in the rush to be seen to buy young, bolster the team's identity, beat PSG and boost the popularity of the president - what was Barcelona's plan for De Jong? How did they expect to improve or even accommodate this talented 21-year-old? In the first game of the season, Barcelona's then-coach Ernesto Valverde offered a show of faith by fielding him ahead of Sergio Busquets. But Barca struggled, Busquets was restored and De Jong shifted to the right of the midfield three. In his two and a half years, De Jong has played right, left, defensive and attacking midfield, as well as right-wing and centre-back. He has played under four permanent coaches, seen Messi leave and Barcelona surrender its status as a serious European club. His best position remains unclear and his future is now in question. Barcelona's financial meltdown means any significant sale is always tempting and De Jong has been linked to both Manchester clubs in recent weeks. He would likely fetch more than 50 million euros in either January or the summer transfer window. For now, the onus is on De Jong to take what could be his last chance under Xavi, to satisfy lofty expectations and make good on a signing that was arguably seen as an end in itself, with too little thought for what came next.

Document ID: 7

Bangladesh suffered an embarrassing defeat by an innings and eight runs against Pakistan in the rain-affected second Test of the two-match Test series at the Sher-e-Bangla National Cricket Stadium in Mirpur. Having crumbled to 87 all out in the morning on the fifth day, the hosts were asked to follow on while trailing by 213 runs. And after 83 overs of batting in the second innings, the Tigers were bundled out for 205. After two 40+ knocks from Liton Das and Mushfiqur rahim in the second innings, Bangladesh at one stage found themselves at 198 for six after 75 overs and were in with a chance of salvaging a draw after Shakib scored a successive team-highest score and more importantly had brought up an impressive fifty partnership with Mehedi Hasan Miraz. But then Pakistan skipper Babar Azam pulled a rabbit out of the hat and decided to bring himself on to bowl the second over of his Test career, in an attempt to break the seventh-wicket stand. The decision immediately paid dividends for the visitors as Miraz was trapped in front trying to play across the line. Miraz's patient 70-ball 14-run knock was spoiled by the sweep shot, which also got the better of him in the first innings. There was still some light at the end of tunnel for the Tigers with Shakib Al Hasan still at the crease. However, he too departed in the next over, on 63 off 130 balls, having failed to read a sharp turning delivery from Sajid Khan which ended up dismantling the off stump. Sajid went on to scalp the wickets of Khaled Mahmud and Tajjilul Islam after Tajjilul and Ebadot Hossain frustrated the visitors with a 34-ball partnership for the last wicket stand. With four wickets in the second innings, Sajid accomplished a stellar 12-wicket haul in his fourth appearance in Tests while Pakistan add maximum ICC World Test Championships points with a 2-0 series win against Bangladesh.

Document ID: 8

Premier allrounder Shakib Al Hasan has achieved another career milestone as he became the quickest player to reach the elite club of players with 4000 runs and 200 wickets in Test cricket. After scoring a team-highest of 33 in the first innings of the Dhaka Test yesterday, Shakib was 34 runs short of the landmark on his 59th appearance for the Tigers in the longest format. And with an inside-out drive to Sajid Khan for a boundary in the 56th over of the second innings, the champion performer finds himself among the illustrious company of the likes of Gary Sobers, Ian Botham, Kapil Dev, Jacques Kallis and Daniel Vettori. The left-hander, currently batting to salvage a draw for his side against the odds on the final session of the Test against Pakistan, took 10 Tests fewer than Botham, who achieved the double feat from 69 appearances. Despite being wicketless in the ongoing Test, Shakib boasts 215 wickets in his tally so far, 20 short of Sobers.

Document ID: 9

Mushfiqur Rahim fell two runs short of a half century as Bangladesh keep spiralling deep into trouble on the fifth day of the Dhaka Test. At Tea, Bangladesh are trembling at 147 for 6 after 52.1 overs, with a session left to play on the fifth day of the second Test at the Sher-e-Bangla National Stadium in Mirpur. When Mushfiqur attempted a reverse sweep to Sajid Khan in the 48th over of the second innings, the writing was on the wall for the experienced batter it seemed. And just at the stroke of Tea, Mushfiqur lost the plot completely when he called for a cheeky single and then managed to ran himself out at the striker's end after his 136-ball 48-run knock. The umpire was reluctant to pass the call to third umpire at first but replays showed that Mushfiqur failed to slide the bat onto the ground by the moment wicketkeeper Mohammad Rizwan dislodged the bails. Shakib Al Hasan is unbeaten on 25 off 56 deliveries and will be hoping to get all the support from the likes of Mehedi Hasan Miraz and Tajjilul Islam. Pakistan, on the other hand, will be looking to wrap up this game and clean sweep the Test series as early as possible, in case bad light stops play prematurely. Shaheen Afridi, Hasan Ali has scalped two wicket for the visitors while Sajid Khan, the star of the first innings with a eight-for, picked the important wicket of Liton Das.

Document ID: 10

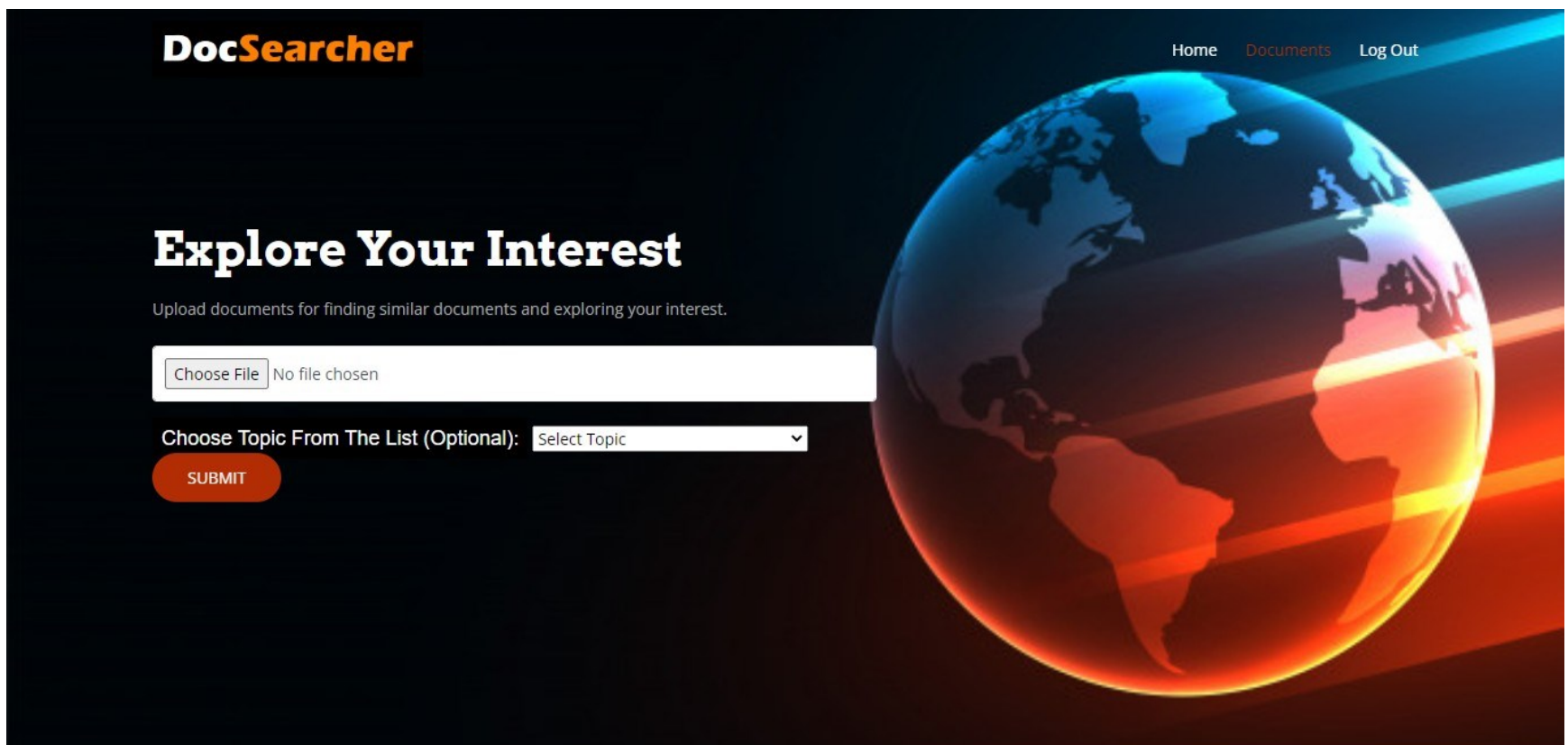
Bangladesh surpassed their first innings tally of 87 to reach three figures in the second innings before Liton Das perished as the hosts trail by 104 runs at the first drinks break of the post-lunch session after being enforced a follow-on by Pakistan on the fifth and final day of the Dhaka Test. The Tigers are in deep water at 109 for five after 38 overs. Shakib Al Hasan, the team-highest scorer of the first innings, has joined Mushfiqur Rahim, on 31 off 97 balls, at the crease. Liton began the post-lunch session by bringing up the fifty partnership with Mushfiqur for the fifth wicket, as he rocked on the backfoot to pull a Sajid Khan loose, short delivery past the short leg. However, the partnership could only last up to 73 runs as a similar delivery and shot ten overs later brought the downfall of Liton, on 45 off 81 deliveries. This time around, a fielder in Fawad Alam was present at square leg to take a straightforward catch, much to the delight of Sajid Khan, who added another victim after his eight-wicket haul in the first innings. Earlier in the session, the wicketkeeper batter got lucky against Shaheen Afridi in the next over when an inside edge from a half-hearted defensive stroke narrowly missed the leg stump. Both Liton and Mushfiqur had played proactively on a wicket which has dried up in the sun and looked in a better shape, comparatively than yesterday, for batters. Still, Pakistan bowlers, especially Afridi, has never looked far from breaking the partnership with his movement, pace and bounce. And more promising signs popped up for the visitors as Sajid got one to bounce awkwardly high for Liton in the middle of the 30th over but Liton got lucky again as the ball landed on the vacant silly point area off his glove. However, it did not require a brilliant delivery nor did the fifth day pitch had any role with the manner Liton got out.

Document ID: 11

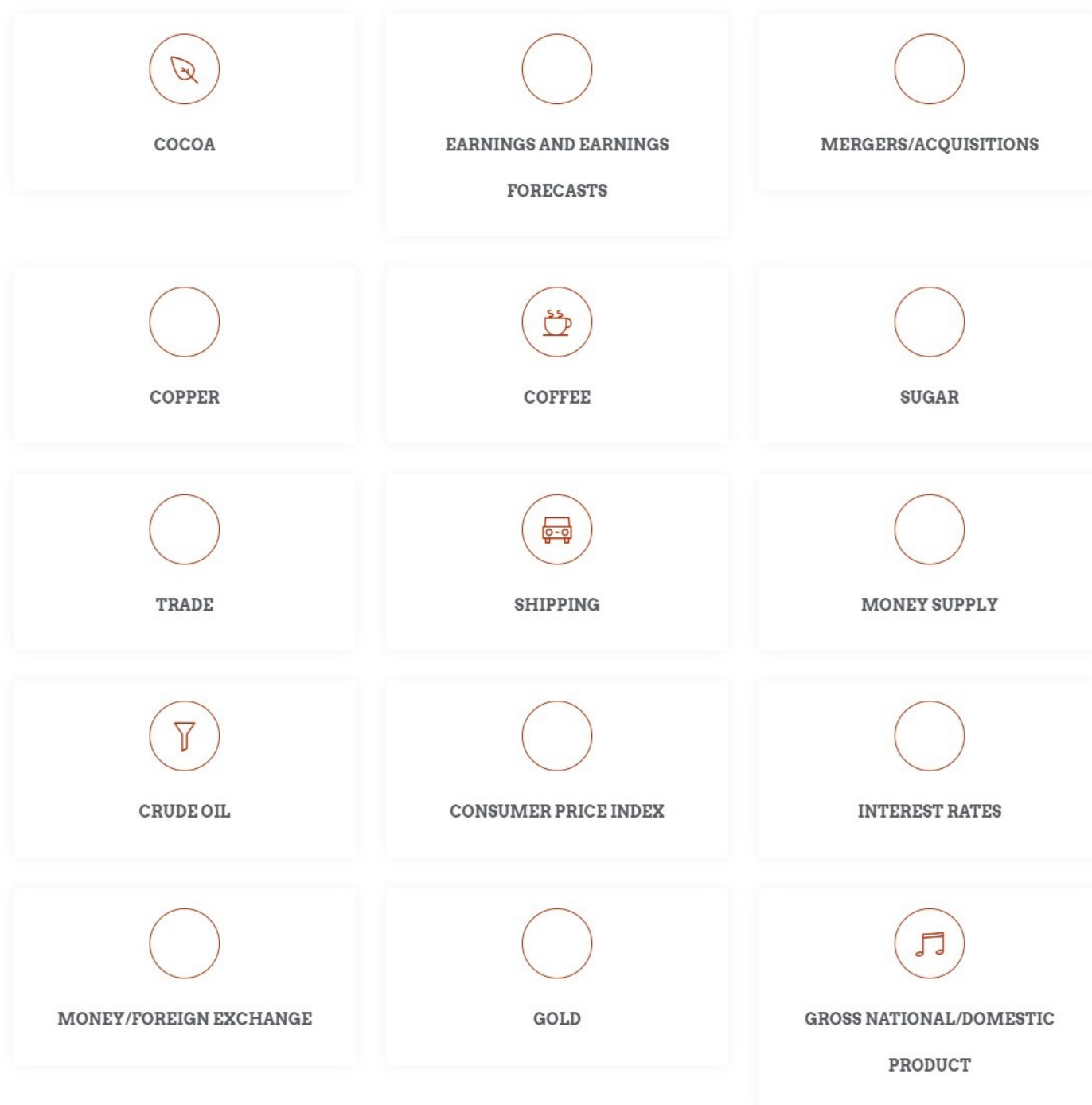
Bangladesh batters Mushfiqur Rahim and Liton Das put on some resistance following the Tigers' top-order collapse after Pakistan had put the hosts to follow-on on the final day of their second Test at the Sher-e-Bangla National Stadium in Mirpur today. Bangladesh were able to score 72 for four before umpires signaled for lunch. However, the Tigers, who are following on, still trail by 141 runs. The duo of Mushfiqur and Liton brought some stability in the middle, stitching together an unbeaten 47-run fifth-wicket stand, which is also the highest total so far by Bangladesh in this game. Mushfiqur remained unbeaten on 16 off 52 deliveries while Liton, who had started very cautiously as he took 21 balls to open his account, notched up a gear after settling in. Liton pounced on the loose deliveries by Pakistan spinners Nauman Ali and Sajid Khan as the right-hander smashed a total of four boundaries during his unbeaten 46-ball 27. Earlier, Pakistan's pace duo of Shaheen Shah Afridi and Hasan Ali ran riot as they took two wickets each to put Bangladesh in a precarious situation. It was the same old struggle of Bangladesh batters against the pace and swing of the Pakistani pacers that saw the Tigers reeling at 25 for four at one stage of the game. Before that, Pakistan had bundled out Bangladesh for 87, the lowest-ever total in Tests in Mirpur, to enforce a follow-on.

Document ID: 12

Bangladesh's fragile top-order was once again exposed by the pace threat of Pakistan as the Tigers lost four wickets inside the first 10 overs after being put to follow-on on the final day of their second Test at the Sher-e-Bangla National Stadium in Mirpur today. After 11 overs, Bangladesh were left reeling at 27 for four with Pakistan pacers Shaheen Shah Afridi and Hasan Ali scalping two wickets each. It was the same old struggle of Bangladesh batters against the pace and swing of the Pakistan pace duo. Debutant Mahmudul Hasan Joy was the first to depart, failing to negotiate a Hasan delivery that nipped in to rattle the right-hander's stumps. Shadman Islam was soon undone by Shaheen as the left-handed batter failed to get his bat down in time and was trapped in front when the namer had one run back inward after having swung the cherry away for the most occasions. Hasan

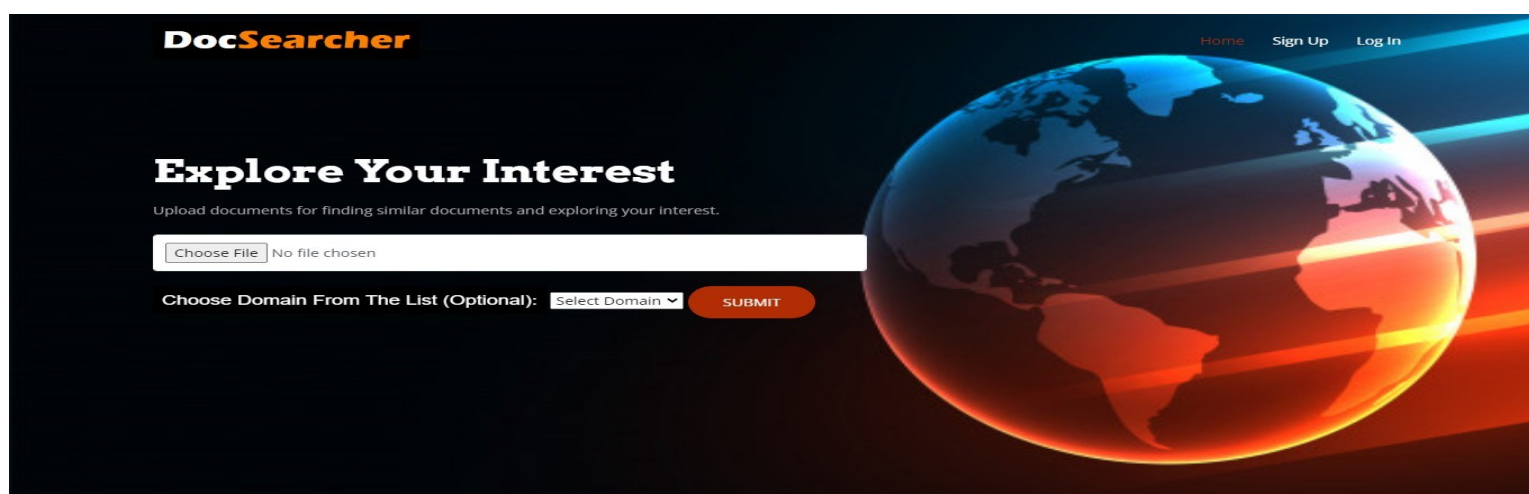


REUTER DOCUMENTS



Reuter

Reuters Dataset repository. Users can view this page. No login or registrations are required. The Reuter repository has fifteen topics containing hundreds of documents on the fifteen issues.



Main Document:

Members of the medical board, formed to find the cause of 11 zebras' deaths, found lead in the grass (fed to the animals) in Bangabandhu Sheikh Mujib Safari Park in Gazipur. This has happened due to air pollution, said Bangladesh Agriculture University Professor Dr Md Abu Hadi Nur Ali Khan, a member of the board. The findings were revealed during the investigation into the deaths of 11 zebras at the Safari Park in Gazipur's Sreepur upazila. He said, "The lead was found during lab test. Adjacent areas of the park have industries and heavy traffic." "Eight out of the 11 zebras died because of air pollution and cold during winter. Reasons behind deaths of the rest could not be confirmed yet," he added. Besides, inbreeding could be another cause, Prof Abu Hadi Nur Ali Khan said. The issues were identified in the investigation report, he said, adding that there is one veterinarian at the park which is inadequate. The lab report has been submitted to the ministry concerned and a meeting will be held between the investigation committee, formed by the ministry, and the medical board tomorrow.

Suggested Domain: environment ,

Similar Documents

Document ID: 214

Domain: environment

Members of the medical board, formed to find the cause of 11 zebras' deaths, found lead in the grass (fed to the animals) in Bangabandhu Sheikh Mujib Safari Park in Gazipur. This has happened due to air pollution, said Bangladesh Agriculture University Professor Dr Md Abu Hadi Nur Ali Khan, a member of the board. The findings were revealed during the investigation into the deaths of 11 zebras at the Safari Park in Gazipur's Sreepur upazila. He said, "The lead was found during lab test. Adjacent areas of the park have industries and heavy traffic." "Eight out of the 11 zebras died because of air pollution and cold during winter. Reasons behind deaths of the rest could not be confirmed yet," he added. Besides, inbreeding could be another cause, Prof Abu Hadi Nur Ali Khan said. The issues were identified in the investigation report, he said, adding that there is one veterinarian at the park which is inadequate. The lab report has been submitted to the ministry concerned and a meeting will be held between the investigation committee, formed by the ministry, and the medical board tomorrow.

Similarity Index: 0.84

Document ID: 232

Domain: environment

An African lioness died today due to sickness at the Bangabandhu Sheikh Mujib Safari Park in Gazipur's Sreepur. Amir Hossain Chowdhury, head of the forest conservators, confirmed the death of the lioness. The 11-year-old lioness first fell ill on August 11 last year. Since then, veterinary doctors from the Safari Park, National Zoo, Bangladesh Agriculture University, Mymensingh Veterinary Teaching Hospital continuously had been giving treatment to her, said Amir. Due to the illness, she had been suffering from various complications including bleeding from the mouth and respiratory problems. Around 4:15 pm yesterday, the lioness started trembling and later Safari Park Surgeon Dr Md Mostafizur Rahman gave her treatment, added Amir. And again today around 7:30 am, she was given treatment as she suffered from breathing problems. Later around 1:00 pm, the lioness was found dead, he said. A postmortem would be conducted on the lioness, Amir also said. At least 11 zebras and a tiger died at the park last month.

Similarity Index: 0.83

Document ID: 230

Domain: environment

Steps will be taken against those involved in negligence that caused deaths of 11 zebras, and a lioness at Bangabandhu Sheikh Mujib Safari Park in Gazipur, Minister for Environment, Forest and Climate Change Md Shahab Uddin said today. Exemplary measures will be taken based on the report by the investigation committee formed by the ministry, he said. The minister came up with the remarks while visiting the safari park in Sreepur upazila, our Gazipur correspondent reports. He expressed hope that the committee would submit its report within 10 working days. The report will reveal the information about the causes of the animals' deaths, irregularities and those responsible, he added. State Minister Habibunnahar, local lawmaker Md Iqbal Hossain Sabuj, Secretary Md Mostafa Kamal, head of forest conservators Md Amir Hossain Chowdhury were present there.

Similarity Index: 0.80

Document ID: 236

Domain: environment

The High Court today criticised the authorities concerned of the government for their failure in taking effective measures to control air pollution in Dhaka and surrounding areas despite its repeated directives. "Public health is at serious risk due to air pollution. The situation is not improving although this court has delivered several directives to this effect. In this situation the court cannot sit idle," the HC bench of Justice Md Ashfaquul Islam and Justice Md Iqbal Kabir said during hearing a writ petition. The bench directed the director general of environment department and deputy commissioners of five districts -- Dhaka, Manikganj, Munshiganj, Narsingdi and Narayanganj -- to be connected with it during next hearing of the petition on February 15. The HC also asked them to submit separate lists of illegal brick kilns before the court on that day. The writ petition was moved by lawyer Manzill Murshid on behalf of Human Rights and Peace for Bangladesh (HRPB). On February 4 last year, the HC asked the DG of Fire Service and Civil Defence to take necessary steps to spray water from its vehicles on urgent basis at the entrances of the capital including Gabotli, Jatrabari, Purbachal, Keraniganj and Tongi.

Similarity Index: 0.80

Document ID: 400

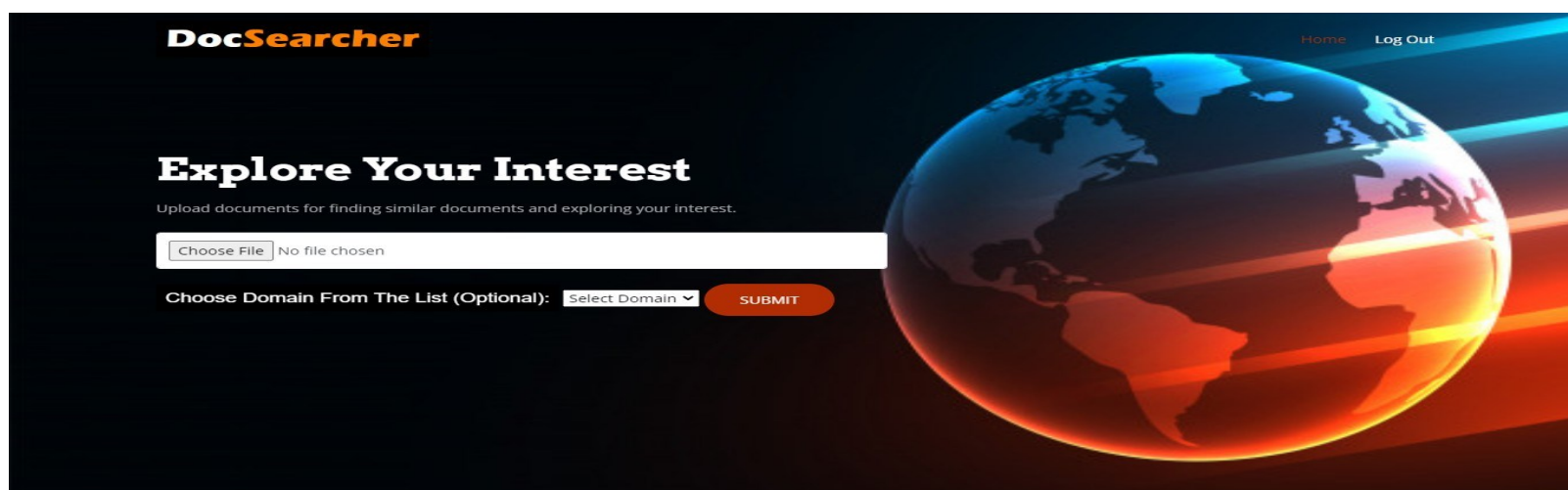
Domain: youth

The authorities of Dhaka University today decided not to hold any first-year honours admission tests for 'Gha' or D unit, under the social science faculty, from the 2021-2022 academic year. The university's deans' sub-committee, at a meeting of DU's general admission committee today at Nabab Nawab Ali Chowdhury Senate Bhaban on the university campus, was assigned the responsibility to form a policy to admit students to Gha unit seats via admission tests for Ka, Kha, Ga (A, B, C) units and Cha unit, reports our DU correspondent. DU Vice-Chancellor Prof M Akhtaruzzaman chaired the meeting. Earlier last year, DU authorities had decided to reduce the burden of entrance exams on the students. All students can get admissions to this unit by taking the entry tests of their respective units (Ka, Kha, Gaz and Cha units), VC Prof Akhtaruzzaman told The Daily Star confirming the development. "We have prior experience as the university admitted students following this system earlier," he added. The DU authorities are also going to redefine the number of seats for admission in the university, considering its capacity and capability alongside fulfilling the national and international requirements in order to improve the quality of education and create adequate skilled human resources. According to DU's public relations office, this agenda will be finalised by the university's Academic Council's meeting immediately. DU academic year 2020-2021 admission tests were held at all divisional cities across the country in October, last year.

Similarity Index: 0.80

Domain Independent Similar Document Search for Non-registered User

One of the developed prototype's two main features is Domain-Independent Similar Document Search for Non-registered User. Users can upload a document to find similar ones via this page. This page shows the user uploaded document, the Suggested domain from the classifiers, ten relevant documents, and their similarity index for this repository. Similar document and domain names are also shown for the user's convenience. This is a sample search result of the domain-independent case where the user will not choose any domain before classifying a document.



Main Document:

Members of the medical board, formed to find the cause of 11 zebras' deaths, found lead in the grass (fed to the animals) in Bangabandhu Sheikh Mujib Safari Park in Gazipur. This has happened due to air pollution, said Bangladesh Agriculture University Professor Dr Md Abu Hadi Nur Ali Khan, a member of the board. The findings were revealed during the investigation into the deaths of 11 zebras at the Safari Park in Gazipur's Sreepur upazila. He said, "The lead was found during lab test. Adjacent areas of the park have industries and heavy traffic." "Eight out of the 11 zebras died because of air pollution and cold during winter. Reasons behind deaths of the rest could not be confirmed yet," he added. Besides, inbreeding could be another cause, Prof Abu Hadi Nur Ali Khan said. The issues were identified in the investigation report, he said, adding that there is one veterinarian at the park which is inadequate. The lab report has been submitted to the ministry concerned and a meeting will be held between the investigation committee, formed by the ministry, and the medical board tomorrow.

Suggested Domain: environment ,

Similar Documents

Document ID: 214

Domain: environment

Members of the medical board, formed to find the cause of 11 zebras' deaths, found lead in the grass (fed to the animals) in Bangabandhu Sheikh Mujib Safari Park in Gazipur. This has happened due to air pollution, said Bangladesh Agriculture University Professor Dr Md Abu Hadi Nur Ali Khan, a member of the board. The findings were revealed during the investigation into the deaths of 11 zebras at the Safari Park in Gazipur's Sreepur upazila. He said, "The lead was found during lab test. Adjacent areas of the park have industries and heavy traffic." "Eight out of the 11 zebras died because of air pollution and cold during winter. Reasons behind deaths of the rest could not be confirmed yet," he added. Besides, inbreeding could be another cause, Prof Abu Hadi Nur Ali Khan said. The issues were identified in the investigation report, he said, adding that there is one veterinarian at the park which is inadequate. The lab report has been submitted to the ministry concerned and a meeting will be held between the investigation committee, formed by the ministry, and the medical board tomorrow.

Similarity Index: 0.84

Document ID: 232

Domain: environment

An African lioness died today due to sickness at the Bangabandhu Sheikh Mujib Safari Park in Gazipur's Sreepur. Amir Hossain Chowdhury, head of the forest conservators, confirmed the death of the lioness. The 11-year-old lioness first fell ill on August 11 last year. Since then, veterinary doctors from the Safari Park, National Zoo, Bangladesh Agriculture University, Mymensingh Veterinary Teaching Hospital continuously had been giving treatment to her, said Amir. Due to the illness, she had been suffering from various complications including bleeding from the mouth and respiratory problems. Around 4:15 pm yesterday, the lioness started trembling and later Safari Park Surgeon Dr Md Mostafizur Rahman gave her treatment, added Amir. And again today around 7:30 am, she was given treatment as she suffered from breathing problems. Later around 1:00 pm, the lioness was found dead, he said. A postmortem would be conducted on the lioness, Amir also said. At least 11 zebras and a tiger died at the park last month.

Similarity Index: 0.83

Document ID: 230

Domain: environment

Steps will be taken against those involved in negligence that caused deaths of 11 zebras, and a lioness at Bangabandhu Sheikh Mujib Safari Park in Gazipur, Minister for Environment, Forest and Climate Change Md Shahab Uddin said today. Exemplary measures will be taken based on the report by the investigation committee formed by the ministry, he said. The minister came up with the remarks while visiting the safari park in Sreepur upazila, our Gazipur correspondent reports. He expressed hope that the committee would submit its report within 10 working days. The report will reveal the information about the causes of the animals' deaths, irregularities and those responsible, he added. State Minister Habibunnahar, local lawmaker Md Iqbal Hossain Sabuj, Secretary Md Mostafa Kamal, head of forest conservators Md Amir Hossain Chowdhury were present there.

Similarity Index: 0.80

Document ID: 236

Domain: environment

The High Court today criticised the authorities concerned of the government for their failure in taking effective measures to control air pollution in Dhaka and surrounding areas despite its repeated directives. "Public health is at serious risk due to air pollution. The situation is not improving although this court has delivered several directives to this effect. In this situation the court cannot sit idle," the HC bench of Justice Md Ashfaqul Islam and Justice Md Iqbal Kabir said during hearing a writ petition. The bench directed the director general of environment department and deputy commissioners of five districts -- Dhaka, Manikganj, Munshiganj, Narsingdi and Narayanganj -- to be connected with it during next hearing of the petition on February 15. The HC also asked them to submit separate lists of illegal brick kilns before the court on that day. The writ petition was moved by lawyer Manzill Murshid on behalf of Human Rights and Peace for Bangladesh (HRPB). On February 4 last year, the HC asked the DG of Fire Service and Civil Defence to take necessary steps to spray water from its vehicles on urgent basis at the entrances of the capital including Gabotli, Jatrabari, Purbachal, Keraniganj and Tongi.

Similarity Index: 0.80

Document ID: 63

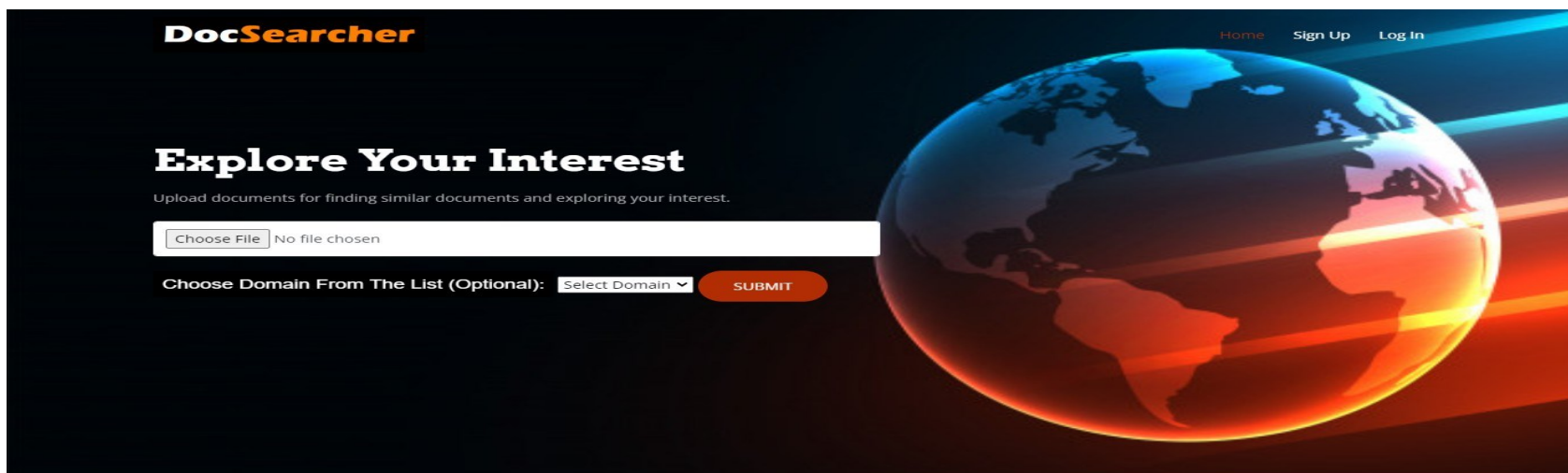
Domain: environment

Citing the constitution that makes it mandatory to protect forest, environment and wildlife, Environment Minister Md Shahab Uddin today said they were not able to conserve wildlife. He came up with the comment at a time when eight elephants were killed last month across Chattogram and Sherpur, mostly by electrocution and shooting by forest grabbers. The minister was addressing a programme titled -- National Result Sharing and Consultation Workshop -- as chief guest, organised by International Union for Conservation of Nature (IUCN) at Bangabandhu International Conference Centre in Dhaka. The conference was organised as part of IUCN project "Feasibility Study of Transboundary Wildlife Corridor in CHT, Chattogram and Cox's Bazar with Myanmar and India Project". Forest Minister Shahab Uddin said according to the constitution's section-18 (kha), the state will conserve its environment, forest and wildlife for present and future generation of the country. "The wildlife is supposed to be in forest. But when forest goes under the clutch of human being, how the wildlife will survive there," he said. He urged other agencies of government to join hand to protect the forest as forest department alone won't be able to do it. Habibun Nahar, deputy minister of the ministry, said, "there is a law in place. But we all know how much of the law is executed on the ground. Md Mostafa Kamal, secretary of the ministry, vowed not to let the giant mammal disappear from Bangladesh territory. "We request IUCN to give us specific data on elephant corridors. We will acquire land for safe movement of elephants if necessary," he said. Monirul H Khan, professor of Zoology at Jahangirnagar University, presented the keynote on the transboundary corridors of elephant while Raquibul Amin presented the overview of the project.

Similarity Index: 0.80

Domain Independent Similar Document Search for Registered User

Registered users can search documents without selecting any particular domain. The feature of user-centric searching is shown in this picture; the registered user can give their preferred domain or topic at the time of registration. Based on their chosen area, the system will show similar documents upon a search.



Main Document:

Members of the medical board, formed to find the cause of 11 zebras' deaths, found lead in the grass (fed to the animals) in Bangabandhu Sheikh Mujib Safari Park in Gazipur. This has happened due to air pollution, said Bangladesh Agriculture University Professor Dr Md Abu Hadi Nur Ali Khan, a member of the board. The findings were revealed during the investigation into the deaths of 11 zebras at the Safari Park in Gazipur's Sreepur upazila. He said, "The lead was found during lab test. Adjacent areas of the park have industries and heavy traffic." "Eight out of the 11 zebras died because of air pollution and cold during winter. Reasons behind deaths of the rest could not be confirmed yet," he added. Besides, inbreeding could be another cause, Prof Abu Hadi Nur Ali Khan said. The issues were identified in the investigation report, he said, adding that there is one veterinarian at the park which is inadequate. The lab report has been submitted to the ministry concerned and a meeting will be held between the investigation committee, formed by the ministry, and the medical board tomorrow.

Suggested Domain: environment ,

Similar Documents

Document ID: 214

Domain: environment

Members of the medical board, formed to find the cause of 11 zebras' deaths, found lead in the grass (fed to the animals) in Bangabandhu Sheikh Mujib Safari Park in Gazipur. This has happened due to air pollution, said Bangladesh Agriculture University Professor Dr Md Abu Hadi Nur Ali Khan, a member of the board. The findings were revealed during the investigation into the deaths of 11 zebras at the Safari Park in Gazipur's Sreepur upazila. He said, "The lead was found during lab test. Adjacent areas of the park have industries and heavy traffic." "Eight out of the 11 zebras died because of air pollution and cold during winter. Reasons behind deaths of the rest could not be confirmed yet," he added. Besides, inbreeding could be another cause, Prof Abu Hadi Nur Ali Khan said. The issues were identified in the investigation report, he said, adding that there is one veterinarian at the park which is inadequate. The lab report has been submitted to the ministry concerned and a meeting will be held between the investigation committee, formed by the ministry, and the medical board tomorrow.

Similarity Index: 0.84

Document ID: 232

Domain: environment

An African lioness died today due to sickness at the Bangabandhu Sheikh Mujib Safari Park in Gazipur's Sreepur. Amir Hossain Chowdhury, head of the forest conservators, confirmed the death of the lioness. The 11-year-old lioness first fell ill on August 11 last year. Since then, veterinary doctors from the Safari Park, National Zoo, Bangladesh Agriculture University, Mymensingh Veterinary Teaching Hospital continuously had been giving treatment to her, said Amir. Due to the illness, she had been suffering from various complications including bleeding from the mouth and respiratory problems. Around 4:15 pm yesterday, the lioness started trembling and later Safari Park Surgeon Dr Md Mostafizur Rahman gave her treatment, added Amir. And again today around 7:30 am, she was given treatment as she suffered from breathing problems. Later around 1:00 pm, the lioness was found dead, he said. A postmortem would be conducted on the lioness, Amir also said. At least 11 zebras and a tiger died at the park last month.

Similarity Index: 0.83

Document ID: 230

Domain: environment

Steps will be taken against those involved in negligence that caused deaths of 11 zebras, and a lioness at Bangabandhu Sheikh Mujib Safari Park in Gazipur, Minister for Environment, Forest and Climate Change Md Shahab Uddin said today. Exemplary measures will be taken based on the report by the investigation committee formed by the ministry, he said. The minister came up with the remarks while visiting the safari park in Sreepur upazila, our Gazipur correspondent reports. He expressed hope that the committee would submit its report within 10 working days. The report will reveal the information about the causes of the animals' deaths, irregularities and those responsible, he added. State Minister Habibunnahar, local lawmaker Md Iqbal Hossain Sabuj, Secretary Md Mostafa Kamal, head of forest conservators Md Amir Hossain Chowdhury were present there.

Similarity Index: 0.80

Document ID: 236

Domain: environment

The High Court today criticised the authorities concerned of the government for their failure in taking effective measures to control air pollution in Dhaka and surrounding areas despite its repeated directives. "Public health is at serious risk due to air pollution. The situation is not improving although this court has delivered several directives to this effect. In this situation the court cannot sit idle," the HC bench of Justice Md Ashfaul Islam and Justice Md Iqbal Kabir said during hearing a writ petition. The bench directed the director general of environment department and deputy commissioners of five districts -- Dhaka, Manikganj, Munshiganj, Narsingdi and Narayanganj -- to be connected with it during next hearing of the petition on February 15. The HC also asked them to submit separate lists of illegal brick kilns before the court on that day. The writ petition was moved by lawyer Manzill Murshid on behalf of Human Rights and Peace for Bangladesh (HRPB). On February 4 last year, the HC asked the DG of Fire Service and Civil Defence to take necessary steps to spray water from its vehicles on urgent basis at the entrances of the capital including Gabotli, Jatrabari, Purbachal, Keraniganj and Tongi.

Similarity Index: 0.80

Document ID: 63

Domain: environment

Citing the constitution that makes it mandatory to protect forest, environment and wildlife, Environment Minister Md Shahab Uddin today said they were not able to conserve wildlife. He came up with the comment at a time when eight elephants were killed last month across Chattogram and Sherpur, mostly by electrocution and shooting by forest grabbers. The minister was addressing a programme titled -- National Result Sharing and Consultation Workshop -- as chief guest, organised by International Union for Conservation of Nature (IUCN) at Bangabandhu International Conference Centre in Dhaka. The conference was organised as part of IUCN project "Feasibility Study of Transboundary Wildlife Corridor in CHT, Chattogram and Cox's Bazar with Myanmar and India Project". Forest Minister Shahab Uddin said according to the constitution's section-18 (kha), the state will conserve its environment, forest and wildlife for present and future generation of the country. "The wildlife is supposed to be in forest. But when forest goes under the clutch of human being, how the wildlife will survive there," he said. He urged other agencies of government to join hand to protect the forest as forest department alone won't be able to do it. Habibun Nahar, deputy minister of the ministry, said, "there is a law in place. But we all know how much of the law is executed on the ground. Md Mostafa Kamal, secretary of the ministry, vowed not to let the giant mammal disappear from Bangladesh territory. "We request IUCN to give us specific data on elephant corridors. We will acquire land for safe movement of elephants if necessary," he said. Monirul H Khan, professor of Zoology at Jahangirnagar University, presented the keynote on the transboundary corridors of elephant while Raquibul Amin presented the overview of the project.

Similarity Index: 0.80

Domain Dependent Similar Document Search

The above-shown picture shows a use case where any user has selected the domain of interest for their search, a similar document along with the suggested domain will be delivered to users.