

An Unsupervised NLP Approach for Cross-lingual Urdu-English Text Summarization: A Framework



By:

Zertashia Shafiq

(Registration No: MS-SE-21-363894)

Supervisor

Dr. Usman Qamar

DEPARTMENT OF COMPUTER & SOFTWARE ENGINEERING,
COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING,
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,
ISLAMABAD

August 06, 2024

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by NS **Zertashia Shafiq** Registration No. 00000363894, of College of E&ME has been vetted by undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the thesis.

Signature : _____

Name of Supervisor: Dr Usman Qamar

Date: 06 AUG 2024

Signature of HOD: _____

(Dr Usman Qamar)

Date: 06 AUG 2024

Signature of Dean: _____

(Brig Dr Nasir Rashid)

Date: 06 AUG 2024

An Unsupervised NLP Approach for Cross-lingual Urdu-English Text Summarization: A Framework

By

Zertashia Shafiq

(Registration No:0000363894)

A thesis submitted to the National University of Science and Technology,

Islamabad

in partial fulfillment of the requirements for the degree of
Master of Sciences in Software Engineering

Thesis Supervisor:

Dr. Usman Qamar

DEPARTMENT OF COMPUTER & SOFTWARE ENGINEERING,
COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING,
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,

ISLAMABAD

August 06, 2024

Dedicated to my exceptional family, friends and teachers whose encouragement and support has been my anchor throughout this journey.

ACKNOWLEDGEMENTS

First, First and foremost, I would like to express my gratitude to Allah Almighty, the most merciful and the most kind, for bestowing His blessings upon me and granting me the strength to complete this work.

I am deeply indebted to my supervisor, **Dr. Usman Qamar**, for his unwavering support and guidance. His insights and expertise have been very invaluable to my research and writing process. Without his continuous support, completing this research would not have been possible.

I would also deeply grateful to my GEC committee members, **Dr. Wasi Haider** and **Sir Jahanzeb** for their valuable feedback and suggestions. Which has greatly improved the quality of my work. I dedicate this to my teachers, family and friends, and thankful to them for their constant support and encouragement.

To my family whose love and encouragement has been a source of motivation throughout this journey.

ABSTRACT

We live in an era where there is linguistic diversity and global interconnectedness. In order to move forward, the ability to bridge language barriers is paramount to facilitate cross-cultural communication. This research study presents a comprehensive exploration of cross lingual Urdu to English extractive text summarization framework using an unsupervised NLP approach. The framework incorporates a sequence of steps using a language specific manually prepared dataset. It integrates text translation, summarization using TextRank algorithm, Rouge score calculation and sentiment analysis to assist seamless language comprehension and conversion.

The motivation behind this research emerges from the vital need to address the linguistic divide in a multilingual society like Pakistan. Here, Urdu serves as a national language, but English also holds a significant importance in various areas especially in a professional and educational background. The primary objective is to develop a framework that will be capable of accurately translating cross lingual content meanwhile preserving a semantic meaning of the context.

The framework involves various components, a manually curated dataset that is paired with human generated summaries, along with rouge score in order to assess the accuracy and effectiveness of the framework-generated summaries.

The methodology encompasses dataset preparation, text translation, summarization, evaluation using rouge scores calculation, and sentiment analysis to give reader a gist of the overall content sentiment. The findings of this study contribute to the advancement of cross lingual text summarization technologies.

Keywords: *unsupervised NLP, machine learning, text summarization, extractive summarization, cross lingual, TextRank, parallel corpus, translation, sentiment analysis, framework*

TABLE OF CONTENTS

By	3
ACKNOWLEDGEMENTS	2
ABSTRACT	3
TABLE OF CONTENTS	4
LIST OF TABLES	8
CHAPTER 1	1
1. INTRODUCTION	1
Motivation & Background	1
Leveraging unsupervised over supervised approaches	1
1.1. Problem Statement.....	2
1.2. Objective.....	3
1.3. Scope	3
1.4. Research Contributions.....	3
1.5. National Needs	3
1.6. Applications.....	4
1.7. Thesis Structure	4
CHAPTER 2:.....	5
2. Important Concepts for Cross-Lingual Summarization Framework.....	5
2.1. Concept of Unsupervised Natural Language Processing NLP	5
2.2 Cross-lingual Text Summarization	6
2.3 ROUGE Evaluation Metrics.....	7
2.4 Extractive Text Summarization	8
2.5 TextRank Unsupervised NLP Algorithm	8
CHAPTER 3:.....	10
3. Literature Review	10
3.1 Research Questions	10
3.2 Objectives of Literature Review.....	10
3.3 Keywords.....	10
3.4 Inclusion/exclusion Criteria.....	11
3.5 Related Work	11
3.6 Analysis.....	16
CHAPTER 4:.....	17
4. Challenges In Cross-Lingual Text Summarization Using Unsupervised NLP	17
4.1 Dataset Availability for Urdu Parallel Corpus	17
4.2 Urdu Linguistic – Semantic Ambiguity	17

4.3	Adaptation & Transfer of Cross-Lingual Context	17
4.4	Linguistic Diversity- Domain Specific Terminology & Jargon	18
4.5	Validation & Evaluation of Generated Results.....	18
Chapter 5:		19
5.	Proposed Cross-Lingual Extractive Text Summarization Framework & Its Working.....	19
5.1	Dataset Preparation.....	19
5.2	Urdu to English Text Translator- Google Translator Library.....	26
5.3	Text Summarization Using TextRank Algorithm	28
5.4	Sentiment Analysis	32
5.5	ROUGE Score Calculation	34
CHAPTER 6:.....		36
6.	Results and Analysis.....	36
Chapter 7:.....		56
7.	Conclusion and Future Work.....	56
7.1	Conclusions	56
7.2	Future Work	57
REFERENCES		58

TABLE OF FIGURES

FIGURE 1 2.2 BASIC OVERVIEW OF CROSS-LINGUAL SUMMARIZATION	6
FIGURE 2 2.3 ROUGE METHOD PROCESS	7
FIGURE 3 5.1.1 URDU TEXT COLUMN OF DATASET	20
FIGURE 4 5.1.2 TRANSLATED ENGLISH TEXT	21
FIGURE 5 5.1.2 REFERENCE ENGLISH SUMMARY	21
FIGURE 6 5.1.3 MANUALLY WRITTEN SUMMARIES	22
FIGURE 7 5.2 TRANSLATION PROCESS USING GOOGLE TRANSLATOR API	27
FIGURE 8 5.2 CODE SNIPPET OF GOOGLE TRANSLATOR	28
FIGURE 9 5.3 TEXT PRE-PROCESSING FLOW CHART	29
FIGURE 10 5.3 TEXTRANK ALGORITHM WORKING FLOW CHART	30
FIGURE 11 5.3 CODE SNIPPET OF TEXTRANK	31
FIGURE 12 5.3 CODE SNIPPET OF TEXTRANK ALGORITHM OUTPUT	31
FIGURE 13 5.4 SENTIMENT ANALYSIS FLOW	33
FIGURE 14 .5 REFERENCE SUMMARY CODE SNIPPET	34
FIGURE 15 5.5 FRAMEWORK-GENERATED SUMMARY CODE SNIPPET	34
FIGURE 16 5.5 ROUGE SCORE FUNCTION CODE SNIPPET	35
FIGURE 17 5.5 CODE SNIPPET ROUGE SCORE FINAL STEP	35
FIGURE 18 6.0 FIRST COLUMN OF DATASET – URDU TEXT CONTENT	37
FIGURE 19 6.0 TRANSLATION CODE SNIPPET	38
FIGURE 20 6.0 TRANSLATION OUTPUT SNIPPET	38
FIGURE 21 6.0 SECOND COLUMN OF DATASET – ENGLISH TRANSLATED TEXT	38
FIGURE 22 6.0 THIRD COLUMN OF DATASET – REFERENCE ENGLISH SUMMARY	39
FIGURE 23 6.0 MANUALLY WRITTEN SUMMARIES COLUMN	39
FIGURE 24 6.0 MANUAL SUMMARIES ROUGE SCORE CALCULATION CODE SNIPPET-1	39
FIGURE 25 6.0 MANUAL SUMMARIES ROUGE SCORE CALCULATION CODE SNIPPET-2	40
FIGURE 26 6.0 MANUAL SUMMARIES ROUGE SCORE CALCULATION OUTPUT SNIPPET	40
FIGURE 27 6.0 ROUGE SCORES OF MANUALLY WRITTEN SUMMARIES	41
FIGURE 28 6.0 TEXTRANK ALGORITHM CODE-1	42
FIGURE 29 6.0 TEXTRANK ALGORITHM CODE-2	42
FIGURE 30 6.0 FRAMEWORK-GENERATED TEXT OUTPUT	42
FIGURE 31 FIGURE 14-6.0 FRAMEWORK-GENERATED TEXT COLUMN	43
FIGURE 32 6.0 CODE SNIPPET FOR SENTIMENT ANALYSIS	44
FIGURE 33 6.0 OUTPUT FOR SENTIMENT ANALYSIS CODE	44
FIGURE 34 6.0 ROUGE SCORE CALCULATION CODE SNIPPET-1	45
FIGURE 35 6.0 ROUGE SCORE CALCULATION CODE SNIPPET-2	45
FIGURE 36 6.0 ROUGE SCORE OUTPUT FOR FRAMEWORK-GENERATED SUMMARY	45
FIGURE 37 6.0 CALCULATED ROUGE SCORES FOR FRAMEWORK-GENERATED SUMMARIES	46
FIGURE 38 6.0 AVERAGE OF EACH ROUGE SCORE FOR MANUAL SUMMARIES	47
FIGURE 39 6.0 AVERAGE OF EACH ROUGE SCORE FOR GENERATED SUMMARIES	47
FIGURE 40 6.0 ROUGE SCORES COMPARISON BAR GRAPH	48
FIGURE 41 6.0 COMPARISON LINE GRAPH FOR ROUGE-1 PRECISION TREND	49
FIGURE 42 6.0 COMPARISON LINE GRAPH FOR ROUGE-1 F1 SCORE TREND	50
FIGURE 43 6.0 COMPARISON LINE GRAPH FOR ROUGE-2 F1 SCORE TREND	50
FIGURE 44 6.0 COMPARISON LINE GRAPH FOR ROUGE-L F1 SCORE TREND	51
FIGURE 45 6.0 PRECISION TREND COMPARISON WITHIN MANUAL SUMMARY ROUGE SCORES	51
FIGURE 46 6.0 PRECISION TREND COMPARISON WITHIN FRAMEWORK SUMMARY ROUGE SCORES	52

FIGURE 47 6.0 RECALL TREND COMPARISON WITHIN MANUAL SUMMARY ROUGE SCORES.....	52
FIGURE 47 6.0 RECALL TREND COMPARISON WITHIN MANUAL SUMMARY ROUGE SCORES.....	53
FIGURE 48 6.0 RECALL TREND COMPARISON WITHIN FRAMEWORK SUMMARY ROUGE SCORES.....	53
FIGURE 49 6.0 F1 TREND COMPARISON WITHIN MANUAL SUMMARY SCORES.....	54
FIGURE 50 6.0 F1 TREND COMPARISON WITHIN FRAMEWORK SUMMARY SCORES.....	54

LIST OF TABLES

TABLE 1 3.5-1 PARALLEL APPROACH RELATED WORK	14
TABLE 2 5.1.4.1 ROUGE SCORE CALCULATION.....	26

CHAPTER 1

1. INTRODUCTION

In an increasingly interconnected world, it is essential to facilitate cross-cultural communication and bridge linguistic barriers. If we talk about a multilingual country like Pakistan, citizens often face challenges in accessing and understanding the information available in English. Since the national language of Pakistan is Urdu, but the official global communication language is English, it can be difficult for people to grasp the context of given text in English language. Keeping in mind these challenges, it is crucial to have frameworks and systems that assist people with cross-lingual text communication. The emergence of cross-lingual text summarization has become a critical area of research. The main purpose is to transform the content of one language, in our case a low resource language, to another language while preserving its essence and meaning. This paper presents a framework for cross-lingual text summarization from Urdu to English using unsupervised NLP. This research study endeavors to contribute to the advancement of cross-lingual text summarization without using a training model framework. It addresses the critical need for an effective cross-lingual summarization framework in a multilingual society like Pakistan. It facilitates cross-cultural understanding in an increasingly interconnected world.

Motivation & Background

The motivation for this research stems from the need to address the linguistic divide in a multilingual country like Pakistan. It enhances the information accessibility for Urdu speaking individuals and communities. In Pakistan, even though Urdu is the national language; but in terms of academic, business and international context, English holds a significant importance.

It is a country, which is rich in linguistic diversity. There are many people who need help in understanding Urdu language especially in international context. On the other hand, many native Urdu-speaking people also finds it difficult to convey their message in English language. To address the issue of coexistence of these languages, an efficient framework is needed to summarize the Urdu text and give the important information in English language to the readers. Facilitating access to information and promoting communication across language boundaries.

Motivation behind using unsupervised NLP was to deal with the intense model training process of a low resource language like URDU. In a supervised trained model, it is difficult to find and made Urdu to English parallel corpus. Not to mention the amount of computational resources required for model training. Instead, this paper will move forward with an unsupervised NLP algorithm incorporated in the framework in order to deal with a low resource Urdu language.

Leveraging unsupervised over supervised approaches

This section of the study explains why we are leveraging unsupervised NLP approach for cross-lingual text summarization instead of using supervised approaches. The choice between two approaches carries a substantial impact on the scalability, efficiency, and adaptability of the framework. Using unsupervised NLP approach like TextRank offers unparalleled adaptability and flexibility, especially dealing with a low resource

language.

Requiring annotated training data can be a challenging task in handling diverse textual data. On the other hand, using supervised approach, basic portion of the framework will rely on labeled examples for training, not to mention the number of computational resources that will be required. Unsupervised algorithm like TextRank autonomously summarize and analyze textual content. This will make it suitable for low resource diverse language like Urdu where labeled corpora may be difficult to get hands on.

Unsupervised approach mitigates the risk of bias and the need for human annotations, since algorithm operates solely on intrinsic properties of the text. Whereas supervised NLP approaches are prone to annotations bias. Which in return can heavily influence the quality and performance capabilities of the model.

Dealing with Urdu to English cross-lingual summarization, the biggest challenges are often related to the quality and availability of a linguistic resources, including parallel corpora, dictionary and language models. Unsupervised methods alleviate this dependency by leveraging generic patterns and features of a low resources' language. We also cannot ignore the availability of computational resources when it comes to training and testing language models in supervised learning. In contrast, unsupervised learning operates efficiently on large-scale datasets involving diverse and voluminous textual content.

Domain adaptability is another factor for choosing unsupervised approach over supervised. Algorithm like TextRank enables the data to adapt impeccably to various domains, genres and topic without having to deal with domain specific training data. Besides that, the transparency and interpretability of TextRank algorithm allow users to interpret the summarization process, especially while dealing with framework formation. Supervised models on the hand exhibits black-box behavior, making it difficult to understand the mechanisms driving their predictions. In conclusion, by leveraging unsupervised model empowers the framework to achieve effective and accurate summarization in diverse linguistic contexts.

1.1. Problem Statement

In a digitally connected world, effectively passing on the information from a low resource language like Urdu to a elevated resource language like English can be a challenging task. With the absence of linguistic resources and domain-specific data, it is difficult to carry on the task with traditional supervised approaches, which often utilize annotated data. Hindering the development of cross-lingual text summarization process. There is an absence of scalability and adaptability when it comes to supervised approaches, especially if the content is in a low resource language. It limits the accessibility of critical information and blocks cross-lingual communication.

This study adapts a framework to handle this situation; it leverages machine learning translation capabilities to bridge the linguistic gap between Urdu and English. Then integration of TextRank algorithm; an unsupervised approach enables the process of summarization, facilitating the generation of concise and coherent summaries. There is a need for framework, which reduces the use of complex computational resources while data training and dealing with a low resource language. Meanwhile effectively dealing with a dataset specifically developed for the evaluation of the framework with the help of human generated summaries.

1.2. Objective

The main objective of this study is to facilitate cross-lingual text summarization involving low resource language and reduce the computational resources by using an unsupervised approach.

- Accuracy
- optimize computational efficiency
- Enhance accessibility of a low resource language
- Develop robust framework
- Bridge linguistic gap

1.3. Scope

The scope of this research extends to the development, optimization, evaluation and application of a cross-lingual text summarization framework using unsupervised NLP technique. We instead of training a model for our low resource language, leveraging the unsupervised technique to reduce computational resources and effort. In our approach instead of directly dealing with a low resource Urdu language we first converted into a high resource English language. This will help in maintaining the accuracy of the content along with availability of resources for a high resource language. It will identify limitations and challenges encountered during the process.

The scope includes validating the effectiveness of generated summary by comparing it to the rouge results of human generated summary in our dataset. It also involves applying TextRank unsupervised algorithm into the framework to atomically generate concise and accurate summaries in English, meanwhile preserving the semantic meaning to ensure high quality output. At the end of the results, the framework includes a semantic analysis part to give a gist of the overall essence of the positivity or negativity of the text. The study will deal with extractive summarization technique; our scope does not include abstractive summarization.

1.4. Research Contributions

Our research contributions are described below

- Development of a novel framework
- Enhanced information accessibility
- Less computational resources consumption with unsupervised model
- Contribution to multi-linguistic research
- Improves efficiency in knowledge extraction

1.5. National Needs

In Pakistan, there exist a significant need for the development and implementation of this framework. This country consists of a unique linguistic landscape. English is used as an official language to communicate with the outside world, meanwhile Urdu is our national language and widely spoken across the country. The co-existence of both languages has underscored a necessity for effective mechanisms like cross lingual text summarization to bridge linguistic barriers. It facilitates in accessing the information across diverse communities. The ability to summarize Urdu text document into English assists in research areas, knowledge acquisition, professional development, better understanding of content. Moreover, it promotes cross-border communication, trade and diplomacy by assisting in interaction with audiences and stakeholders. By leveraging the linguistic heritage, Pakistan can embrace the opportunities of digital age and effectively

contribute to the global exchange of ideas and information.

1.6. Applications

There are many possible real world applications of our proposed framework for cross-lingual text summarization

- News aggregation and summarization
- Disseminating information across language barriers, enabling multilingual content curation
- Cross cultural communication
- Enhanced global communication
- Relevant information access for professional and researchers
- Legal and regulatory compliance
- Market analysis; gaining insight into Urdu speaking consumer feedback
- Analyzing and summarizing business reports
- Educational content understanding
- Social media trend monitoring and analyzing for Urdu speaking communities
- Cross lingual information retrieval from web sources
- Understanding Urdu based Government and public policy analysis

1.7. Thesis Structure

Our overall thesis structure is as follows:

Chapter 2: Introduction to main concepts and terminology of framework. In this chapter we will discuss major concepts involved in our framework.

Chapter 3: Literature review. Chapter 3 discusses the previous work done on text summarization models

Chapter 4: Challenges in cross-lingual text summarization process. In this chapter, we will discuss the challenges faced in cross-lingual text summarization with low resource language.

Chapter 5: Proposed Framework and its working. This chapter explained the inner working of our proposed framework and its parts.

Chapter 6: Results and analysis. In this chapter, we applied our framework to our manually prepared dataset and evaluate the results using rouge evaluation. Present the conclusions and evaluate the efficiency of the framework.

Chapter 7: Conclusion and future work. This chapter concludes the whole study and future direction of this research study.

CHAPTER 2:

2. Important Concepts for Cross-Lingual Summarization Framework

In this chapter, we will discuss major concepts and terminologies involved in the development of cross-lingual text summarization framework.

2.1. Concept of Unsupervised Natural Language Processing NLP

The Natural Language Processing NLP is a branch of computational linguistic and artificial intelligence. Unsupervised NLP means that it will focus on processing and understanding language text without relying on labeled data. It does not require a trained label or even explicit data. It leverages structures and patterns within the data, unlike supervised learning approaches, which requires labeled datasets for training. Unsupervised approach derives insights and identify relationships between nodes and edges. Unsupervised NLP targets the hidden patterns, relationships and structures within the given text. The text data does not need to be labeled by humans. It is especially helpful when we do not have labeled data for a low resource language like Urdu.

2.1.1 Key components of unsupervised NLP

There are four major key components of unsupervised NLP

1. Text representation
2. Feature extraction
3. Clustering & classification
4. Dimensionally reduction techniques

There are four major key components of unsupervised NLP. The first one is *Text Representation*. In this, the unsupervised NLP algorithm represents the text data in an organized form that is appropriate for text processing and analysis.

Then the second one is *Feature Extraction*, which includes identifying salient features and patterns within the text. This includes techniques like word embedding, term frequency inverse document frequency TF-IDF, and latent semantic analysis LSA.

The next component is *Clustering & Classification*, which is involve in grouping similar documents or works together based on their context and features.it includes common clustering algorithms like k-means, DBSCAN and hierarchical clustering.

The fourth one is *Dimensionally Reduction techniques*, which deals with high dimensional data. Techniques

like t-distributed stochastic neighbor embedding T-SNE are used.

In text summarization unsupervised technique aims to extract the most significant information from the given text. Our proposed framework uses one of such algorithms called TextRank. It identifies the key sentences based on their score importance. Evaluation of unsupervised NLP can become challenging due to lack of subjective nature of text analysis. Despite its challenges, it remains a powerful technique in understanding, analyzing and deriving value from natural language text.

2.2 Cross-lingual Text Summarization

Cross-lingual text summarization is a process of condensing textual content written in one language into a summarized short form in the target language. The key information and core meaning should be preserved at the end. The end purpose is to enable the users to grasp the whole concept of text in a short summary in a language that they understand.

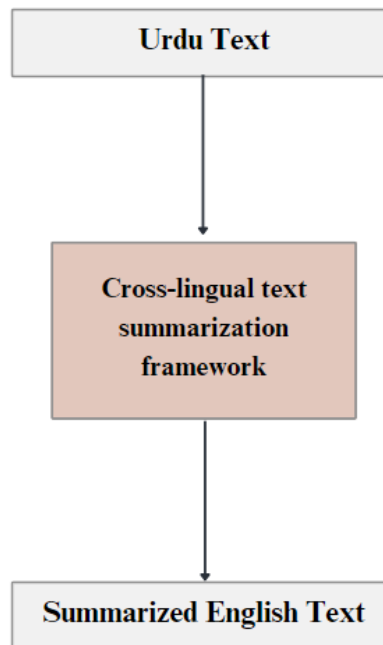


Figure 1 2.2 Basic overview of cross-lingual summarization

The main challenge that occurs in cross lingual text summarization is language variability. This variation includes syntax, structure and semantics of the low resource language we are working with. The next challenge will be to preserve the cultural nuance of the language and accuracy of the summary. In our case, since we are dealing with a low resource language like Urdu, so there might be an ambiguity in language that is the main reason behind using translator to first convert our low resource language into a high resource language. Another limitation is the availability of parallel corpora.

2.3 ROUGE Evaluation Metrics

Rouge evaluation metric, which stands for Recall-Oriented Understudy for Gisting Evaluation, is majorly used to evaluate the quantitative quality of machine learning outputs and text summarization. These metrics evaluates the summaries or translated texts by comparing them to the reference summaries. It provides quantitative measures of the similarities and effectiveness of the summaries. These metrics are essential for evaluating the performance of summarization systems. There are a few important components of rouge metric.

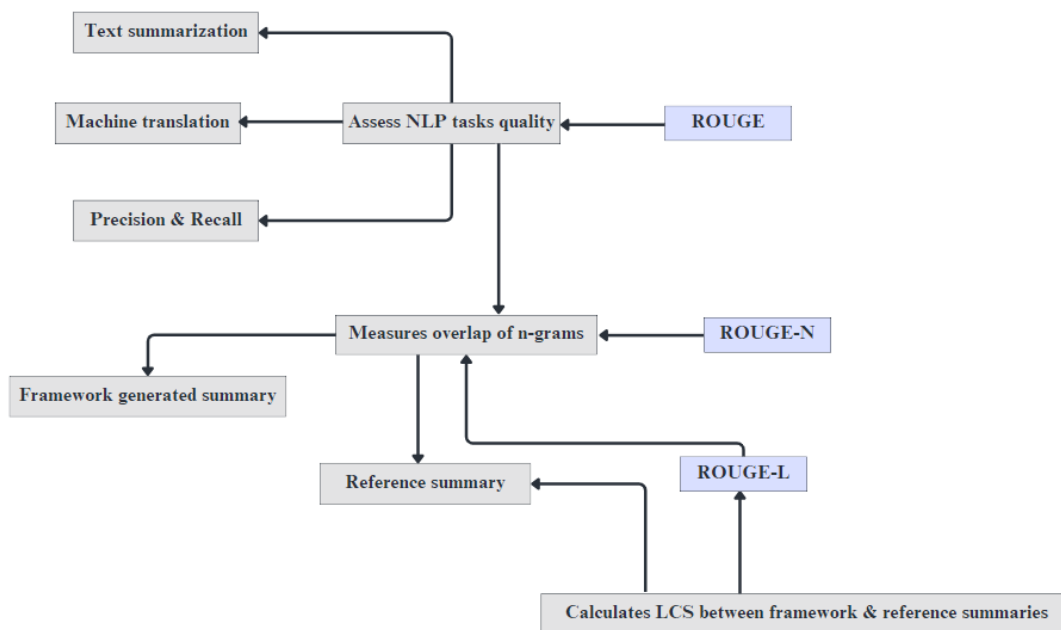


Figure 2 2.3 ROUGE method process

2.3.1 Recall

Recall measures the extent to which the information of the text is being covered by the generated summary, when compared to the reference summary. It is computed by comparing the ratio of number of overlapping words or sequences of words between reference and generated summaries to the total number of n-grams in the reference summary. It is essential since it captures the comprehensiveness that our generated summary show in apprehending the important information from the reference text.

2.3.2 Precision

Precision measures the degree of containing only relevant information from reference summary in the generated summary. It ensures that the generated summary contains the relevant important information and is focused on the main idea along with a concise output.

2.3.3 F1 Score

F1 score is widely used metric for evaluating the complete effectiveness of the translation and summarization systems. It considers both the relevance and coverage of the generated summary. An effectively balanced summary will show a high F1 score.

2.3.4 ROUGE-N (N-gram Overlap)

It evaluates the overlaps of words and sequence of words between the reference and generated summaries. It includes different values of N, such as Rouge-1, Rouge-2 and Rouge-L calculates the rouge metric for longest common subsequence overlap. Rouge-N is responsible for capturing the accuracy of similarity between both summaries.

2.3.5 ROUGE-L (Longest Common Subsequent)

Rouge-L measures the lengthiest common subsequence between the both summary. It takes into account the longest similar sequence of words that is present in both the summaries. It gives an insight into the structural similarity of the output. It is useful for evaluating the coherence and fluency of summaries.

Higher recall values show that there is a strong coverage of information between the generated and reference summaries. Meanwhile higher precision values indicate the more relevancy of information between generated and reference summaries. Therefore, in conclusion, rouge metric plays a crucial role in giving a quantitative evaluation of a subjective generated summarized text.

2.4 Extractive Text Summarization

In extractive text summarization, only the important sentence from the original content is selected and extracted into the summary output. Extractive summarization preserves the original content and context of the provided text. The main aim is to condense a large volume of text into a short more understandable version meanwhile maintaining the original context of the text.

Extractive text summarization happens in various steps. Firstly, the text is tokenized; stop words are removed, along with stemming and other pre-processing techniques. Then sentences are ranked based on their importance and finally all the top ranked sentences formulate the output-summarized text. For quantitatively evaluating the generated summary, we use rouge metrics.

2.5 TextRank Unsupervised NLP Algorithm

TextRank is an unsupervised NLP algorithm that we will be using in our framework for text summarization. Beside text summarization, it is also used for keyword extraction and information retrieval. It draws inspiration from Google's PageRank algorithm. It is graph-based approach, in which the algorithm treats document as a graph, where words or sentences are represented as nodes. In addition, relationship between these nodes is represented as edges. The semantic similarity and relationships between parts of text is capture by the graph structure of this algorithm.

By leveraging metrics like cosine similarity, and Jaccard similarity it calculates the similarity between sentences. These similarity metrics are used for evaluating the strength of the connections between sentences in the graph. In the graphical representation, sentences are represented as nodes, and weight associated with edges shows the similarity between sentences. The more the weight is the stronger is the connection in graph and similarity between sentences.

This algorithm assigns important scores to sentences using an iterative ranking approach. The importance score is based on the connection of a sentence with other sentences in the graph. Therefore, in conclusion the importance score of one sentence will depend on the score of the sentences it is linked with. However, these importance scores are not definitive; they are iteratively getting update based on the importance of the next linked sentence. This iterative process comes to an end once the convergence is reached. The convergence happens when the importance scores stabilize, and no further significant changes happens afterward between the iterations. After that, the sentences with higher scores are considered to be more relevant and will be most likely to be included in the final summary. These key sentences with high importance score contain the salient information in the document. The selected output sentences will capture the key information and ideas of the original text.

The main advantage of TextRank algorithm is that it does not require a labeled dataset, which is very useful for us since we are dealing with a low resource Urdu language. It leverages inherent patterns and relationships to identify the key information and ideas within the text.

CHAPTER 3:

3. Literature Review

In this chapter, we will discuss the previous work done on the cross-lingual text summarization domain. In this way, we can identify the areas in which there is a need of our proposed framework.

3.1 Research Questions

This section will state the important research questions we asked before starting the literature review.

- What other methodologies are available for cross-lingual text summarization?
- How do supervised and unsupervised learning approach effect the summarization process with a low resource language?
- What evaluations metrics are used for evaluating the cross-lingual text summarization process?
- What are the key challenges and limitations associated with cross-lingual text summarization using unsupervised NLP?
- What language specific Urdu to English text summarization parallel corpus available in any previous research?
- What algorithms are used for cross-lingual text summarization and their accuracy level?

3.2 Objectives of Literature Review

- Find research gaps in the current literature
- Identify possible future research directions
- Present our study and research in an organized way
- Survey the literature in order to find answers for our research questions
- Identify which algorithms and evaluation methods are commonly used with more accuracy

3.3 Keywords

Following are the keywords for our literature review

- Cross-lingual text summarization
- Unsupervised NLP
- Translation

- Sentiment analysis
- TextRank
- Linguistic diversity
- Machine learning
- Natural language processing
- Evaluation metrics
- Dataset curation
- Rouge evaluation
- Extractive summarization

3.4 Inclusion/exclusion Criteria

Following are the criteria for our research

- Irrelevant research excluded
- Literature review for papers involving the keywords
- Papers from latest years instead of the old ones
- Papers from the famous database were targeted

3.5 Related Work

In [1] methodologies are used to alleviate the dependency on labeled data by taking advantage of inherent linguistic patterns and cross-lingual similarities. This paper includes a promising direction, the integration of neural network architecture. This architecture includes models such as Transformer model that was proposed by Vaswani et al.(2017). These Transformer models capture the long-range dependencies while enables proficient encoding and decoding of inputs. This model comprises of decoder and encoder layers. Self-attention techniques enable the model to emphasis on relevant phrases and words. Additionally, the incorporation of multi-head attention mechanisms enables the model to attend to multiple representation subspaces simultaneously, further enhancing its expressive power. This paper adopts such approach that integrate encoder-decoder attention techniques. They made experimental evaluation of cross lingual English to Chinese text. Their technique has shown substantial enhancements over baseline approaches.

In [2] the authors have proposed a joint learning method in order to align and summarize for cross-lingual text summarization. It includes training a model to address challenges of previous cross-lingual text summarization. The trained model summarizes text in one language while aligning representations of that text in another language. Their model uses mappers, it combines cross-lingual summarization and monolingual summarization tasks.

In [3] an approach is introduced in order to avoid the error accumulations that were faced by using sequential translation or summarization methods. They used mixed-lingual pre-training method but instead of task specific components, integrated cross-lingual and monolingual tasks. They have improved the cross-lingual summarization by using vast monolingual data for language modeling.

In [4] the paper addresses the shortages in semantic alignment and information compression for low-resource language. This paper uses two stage fine-tuning method for low resource language for cross-lingual summarization. It integrates mPTMs' effectiveness along with pipeline methods' intuitiveness. It leverages transformer-based architecture to achieve high performance.

In [5] this study explores the application of cross-lingual transfer learning. It focuses on two major techniques: VecMap and BiVec. The VecMap technique uses linear transformation to map the generated word embedding by aligning embedding spaces from different languages into the shared vector space. On the other hand, BiVec extracts word representations from multiple languages concurrently, leveraging parallel corpora that generates bilingual word embedding.

In [6] the authors highlighted the evolution of text summarization techniques, its journey from the beginning of the summarization techniques to the contemporary era of natural language processing. This paper discusses notable techniques using machine learning linguistic analysis and semantic understanding to understand and extract key information from the document to formulate a summarized version.

In [7] Natural Language Processing techniques are used for text summarization. These techniques further used the readily available packages in python, R in the programming ecosystem. This paper used two unsupervised techniques TextRank and cosine similarity. Cosine similarity helps in the identification of key phrases and sentences within the text corpus. Whereas TextRank selects the sentences based on their importance level to deliver the context of the text.

In [8] the paper introduces an unsupervised extractive text summarization approach. It uses deep auto-encoders. The framework Ensemble Noisy Auto Encoder ENAE represents a novel advancement in this experiment. It used noise and ensemble aggregation.

In [9] authors address the challenges of text summarization especially in the area of sports specifically in a low resource language like Tamil. The paper introduces a technique where it uses a generative stochastic artificial neural network and integrates it with Natural Language Processing NLP. It increases the identification and abstraction of relevant sentences by using the Restricted Boltzmann Machine RBM and feature vector matrix.

In [10] authors used algorithms like Mathematical Regression MR and Genetic Algorithm GA. In this approach the model generates extractive summaries by learning suitable combinations of feature weights. English religious article dataset is used in this paper to perform experiment.

In [11] the paper discusses trainable machine learning algorithms, and a comparison between their final results for text summarization. The proposed approach utilizes classifiers and comprehensive frameworks like Naïve Bayes, C4.5 decision trees. Uses statistical measures like linguistic attributes and Mean-TF-ISF. The computational results and comparison demonstrate that out of all the used approaches Naïve Bayes classifier demonstrate a better output for text summarization.

In [12] the study focuses on paragraph-level extraction and algorithms used for it. It compares the automated extracts with the human generated summaries. The approaches like statistical analysis of word occurrence and Heuristics-based paragraph extraction are utilized to achieve the results in this paper

In [13] authors have utilized a sentence clustering technique. This approach clusters sentences in the given text based on semantic distance. It then calculates accumulative sentence similarity based on each cluster. It also incorporates multi-feature combination methods, along with capturing underlying semantic structure of the text.

In [14] the study leverages pre-trained language models like BERT, BART. These models are then fine-tuned specifically for Arabic language in this study. It utilizes a novel cross lingual transfer approach to achieve the desire text summarization output. It aims to increase the accuracy and performance of language specific text summarization systems.

In [15] authors have introduced cross-lingual text summarization using a X-SCITLDR dataset for multilingual summarization of CT-TLDR scientific articles. The study includes direct cross-lingual models along with two-stage pipelines model. It also explores the effectiveness of fine-tuning and knowledge distillation models.

Table 1 3.5-1 Parallel Approach Related Work

Paper	Technique and Technology	Improvements	Limitations
[1]	Transformer model, encoder decoder layer	substantial enhancements over baseline approaches	Low resource language dataset
[2]	Cross lingual mapping of context representations + joint training with monolingual summarization	Improved cross-lingual summarization in both supervised and unsupervised settings	Pre-trained monolingual summarizers and additional mappers
[3]	Mixed-lingual pre-training, optimized encoder-decoder architecture	Enhances cross-lingual summarization, by leveraging immense monolingual data accomplishing significant performance boost	Limitation of labeled cross-lingual data for low resource languages
[4]	Two stage fine tuning method for low resource cross-lingual text summarization (TFLCLS)	Efficiently improves information compression and semantic orientation	Availability of low-resource languages dataset
[5]	Cross-lingual transfer learning; VecMap and BiVec	Proposes improvements to VecMap by introducing shared spaces and aligning unshared vocabularies between languages.	Obtaining parallel corpora and aligning diverse linguistic structures
[6]	Transformer-based Models (T5-Small, DistilBART, mBART), deep learning models, machine translation, abstractive summarization	MultiNews and XSum datasets demonstrates significant improvements in the quality of the generated summaries compared to baseline models.	The approach faces limitations such as the scarcity of CLS datasets for English to Hindi, leading to challenges in comprehensive model training

[7]	NLP approach; TextRank, cosine similarity	Improved accuracy and coherence in generating summaries	Limitation in capturing contextual relevance and semantic nuances
[8]	Ensemble Noisy Auto-Encoder (ENAE). Deep auto encoder (AE)	Improves robustness and discriminative power of summarization process	Multi-lingual documents, diverse domains and datasets
[9]	Restricted Boltzmann Machine RBM, feature vector matrix	Semantic relevance and generated summaries coherence, enhances decision-making tasks	Reliance on predefined features, low resource language dataset, domains
[10]	BART, Pegasus, CLSTK, LSTM, co Rank model	Enhances the summarization by creating a robust training dataset	Dependency on the quality and comprehensiveness of the training dataset.
[11]	Monolingual summarization models, Linear mappings	Improves the alignment of contextual information, generate high-quality summaries	accurately aligning and summarizing diverse linguistic structures
[12]	Masked Language Model (MLM), Cross-lingual Masked Language Model (CMLM), Denoising Autoencoder (DAE)	The model benefits from massive unlabeled data, improving its language modeling and cross-lingual representation capabilities.	The reliance on pre-trained models may introduce biases present in the training data
[13]	Sentence clustering based summarization approach	Enhances precision and relevance of text	Complexity of clustering process, computational challenges with large-scale dataset
[14]	Multilingual BERT, BART-50	Summarization quality, paving a way for language specific text summarization	Reliable dataset for abstractive leads, sizeable dataset for model training
[15]	Knowledge distillation models, fine-tuning techniques, multilingual encoder-decoder architecture	Enabling wider accessibility for resources across different domains and languages	Complexity of code-switched texts, diverse linguistic and text contexts

3.6 Analysis

After completing the literature review following are the major gaps

- Cross-lingual summarization with low resource language
- Summarization using unsupervised approach
- Use of computational resources

CHAPTER 4:

4. Challenges In Cross-Lingual Text Summarization Using Unsupervised NLP

In this chapter, we will discuss the challenges faced in cross-lingual text summarization using unsupervised NLP. The framework proposed to tackle these challenges will be discussed in next chapter [5].

4.1 Dataset Availability for Urdu Parallel Corpus

In cross-lingual text summarization, the pivotal point was to find a dataset that meet our requirements for Urdu to English cross-lingual text summarization. It serves as the building block of the whole framework. The process involves sourcing, curating and annotating of the dataset.

One of the major roadblocks in preparing this particular dataset was the scarcity of annotated Urdu Corpora. Urdu is a low resource language, unlikely widely spoken languages like our target language, English. However, for unsupervised approach annotated data serves very little purpose. There are many layers to the Urdu language complexity itself; first, it exhibits significant linguistic diversity across different regions and dialects. Another layer of complexity is added with the domain specificity of the given Urdu textual content. Texts from various domains; sports, health, entertainment, academic, politics, social media and a few more. These types of categorized texts may require a domain-specific preprocessing technique to ensure the accuracy and relevancy of text summarization.

Addressing the cultural and linguistic nuance is one of the significant factors that needs to be fulfilled while preparing the dataset. In addition to preparing dataset, the integration of machine learning methods is essential for the preparation. In conclusion, overcoming the challenges associated with dataset preparation; domain specificity, dataset scarcity, linguistic variations and cultural nuances I essential for the development of a reliable dataset for text summarization framework.

4.2 Urdu Linguistic – Semantic Ambiguity

In the context of Urdu to English text summarization, semantic ambiguity refers to an occurrence where words and phrases have different meanings and interpretations. These complexities pose a challenge to accurately capture the semantic meaning of a low resource language. Especially in Urdu language where certain words or phrases can lead to confusion by having the same word with different meaning depending on the context.

4.3 Adaptation & Transfer of Cross-Lingual Context

The effective transfer & adaptation of context of semantic structures, linguistic patterns, and domain-specific knowledge is a very crucial subject in summarization tasks. There is a lot of linguistic variation between Urdu and English language as our source and target language respectively. There is a significant variation in terms of vocabulary, syntax and morphology of these two languages. Therefore, there is a challenge for direct transfer of content between these two languages. In addition to this, jargon and domain-specific terminology presents additional challenges in cross-lingual transfer. In conclusion, all these mentioned issues need a

nuances approach that will address all these challenges effectively and accurately.

4.4Linguistic Diversity- Domain Specific Terminology & Jargon

There is a direct challenge faced in cross lingual text summarization and that is of lack of direct equivalents of Urdu domain-specific words in English. Like any other diverse and culturally enrich language; Urdu also encompasses a vast array of vocabulary and terminology that is domain specific. Therefore, the crucial task is to accurately understand the context of this domain-specific content. Effective handling of domain-specific terminologies demands access to glossaries, ontologies and knowledge bases relevant to the domain of interest.

4.5Validation & Evaluation of Generated Results

Evaluating and validating result of any developed framework is a crucial step to identify the success level of the system. In case of cross lingual text summarization, since the results are in subjective form, so evaluation metrics should be selected in order to measure the performance of the framework. Nevertheless, the challenging task is to directly be applying any evaluation metrics to the text summarization task. It can show a low accuracy and effectiveness due to the mismatch in reference and generated summaries. The lack of interpretability and subjectivity makes it a challenging task to determine the true quality of the generate summary.

Chapter 5:

5. Proposed Cross-Lingual Extractive Text Summarization Framework & Its Working

In this chapter, we will introduce our framework for cross lingual text summarization using unsupervised NLP. The main purpose of introducing this framework is to deal with all the mentioned challenges in chapter 4. In order to do so, we are leveraging knowledge from a high resource language (ENGLISH) to improve summarization quality in a low resource language that will be URDU in this case. We will be utilizing unsupervised NLP algorithm for extractive text summarization.

Upon completion, this framework aims to address the challenges associated with cross lingual text summarization involving a low resource language. All of this will be done while maintaining the fidelity of original content. This framework approach involves a comprehensive pipeline that encompasses data set preparation, translation, algorithmic implementation, and evaluation and post summarization semantic analysis.

5.1 Dataset Preparation

The cornerstone of this methodology lies in the manual preparation of this dataset. Since we are dealing with a low resource URDU language it is difficult to find the exact dataset that will render to our framework's requirements. This dataset is a curation of URDU corpus texts with their English summaries.

This dataset serves as the foundation of evaluating our framework based on unsupervised NLP algorithm. Each text in the dataset undergoes manual summarization by human. Providing reference summaries and rouge scores against which the framework generated summaries will compare.

Following are the steps that we performed for dataset preparation.

5.1.1 Urdu Corpus News Dataset Creation

We began by curating Urdu corpus dataset, this dataset comprises of a diverse and wide range of topics and domains. The drive behind selecting news data from diverse range of topics is to ensure that our framework works on all sort of data. These diverse fields may include technology news articles, business related news, health, politics, sports and a few other domains.

B	
Translated English text	
<p>Levy points for the White Bread Prize.</p> <p>Novelist Andrea Levy is favorite to win the prestigious White Bread Prize Book of the Year award after winning Novel of the Year with her book <i>Small Island</i>. The book has already won the Orange Prize for Fiction, and is now 5/4 favorite for the £25,000 Whitbread. Another favorite is the biography of Mary Queen of Scots, by John Guy. A panel of judges including Sir Trevor Macdonald, actor Hugh Grant and writer Joanne Harris will decide the overall winner on Tuesday. The five writers in line for the award won their respective categories - debut novel, novel, biography, poetry and children's book - on January 6. <i>Small Island</i>, Levy's fourth novel, is set in post-war London and centers on a landlord and his inhabitants'. One of them is a Jamaican who joins the British army to fight Hitler, but when he settles in Britain, life is difficult because of the uniform. "What could have been a practical or preachy prospect is hilarious, movingly human and eye-popping. It's hard to imagine anyone not enjoying it," the judges wrote. The judges called Guy's <i>My Heart of My Own: The Life of Mary Queen of Scots</i> "an impressive and readable piece of scholarship, which cannot fail to leave the reader moved and intrigued by this most tragic and endearing Queen." Gives." Guy has published many histories, including one on Tudor England. He is a Fellow at Clare College, Cambridge, and became Honorary Research Professor at the University of St Andrews in 2003. Other contenders include Susan Fletcher for <i>Eve Green</i>, who won the first novel prize; Fletcher recently graduated from the University of East Anglia's creative writing course. The fourth book in progress is <i>Corpus</i>, the fourth collection of poems by Michael Simmons Roberts. In addition to writing poetry, Simmons Roberts also makes documentaries. Geraldine McCaughrean is the final contender, winning the children's fiction category for the third time for <i>Not the End of the World</i>. McCaughrean, who went into magazine publishing after studying teaching, previously won the category with <i>A Little Lower than Angels</i> in 1987 and <i>Gold Dust</i> in 1994.</p> <p>The US stock market watchdog's chairman has said he is willing to soften tough new US corporate governance rules to ease the burden on foreign firms. In a speech at the London School of Economics, William Donaldson promised "several initiatives". European firms have protested that US laws introduced after the Enron scandal make Wall Street listings too costly. The US regulator said foreign firms may get extra time to comply with a key clause in the Sarbanes-Oxley Act. The Act comes into force in mid-2005. It obliges all firms with US stock market listings to make declarations, which, critics say, will add substantially to the cost of preparing their annual accounts. Firms that break the new law could face huge fines, while senior executives risk jail terms of up to 20 years. Mr Donaldson said that although the Act does not provide exemptions for foreign firms, the Securities and Exchange Commission (SEC) would "continue to be sensitive to the need to accommodate foreign structures and requirements". There are few, if any, who disagree with the intentions of the Act, which obliges chief executives to sign a statement taking responsibility for the accuracy of the accounts. But European firms with secondary listings in New York have objected - arguing that the compliance costs outweigh the benefits of a dual listing. The Act also applies to firms with more than 300 US shareholders, a situation many firms without US listings could find themselves in. The 300-shareholder threshold has drawn anger as it effectively blocks the most obvious remedy, a delisting. Mr Donaldson said the SEC would "consider whether there should be a new approach to the deregistration process" for foreign firms unwilling to meet US requirements. "We should seek a solution that will preserve investor protections" without turning the US market into "one with no exit", he said. He revealed that his staff were already weighing up the merits of delaying the implementation of the Act's least popular measure - Section 404 - for foreign firms. Seen as particularly costly to implement, Section 404 obliges chief executives to take responsibility for the firm's internal controls by signing a compliance statement in the annual accounts. The</p>	

Figure 4 5.1.2 Translated English Text

These English summaries will serve as the reference summaries that then further use for evaluating rogue scores against both human generated summaries and framework-generated summaries.

C	
Reference English Summary	
Lev	The five writers in line for the award won their respective categories - first novel, novel, biography, poetry and children's book - on 6 January. The book has already won the Orange Prize for fiction, and is now 5/4 favourite for the £25,000 Whitbread. Novelist Andrea Levy is favourite to win the main Whitbread Prize book of the year award, after winning novel of the year with her book <i>Small Island</i> . The other contenders include Susan Fletcher for <i>Eve Green</i> , which won the first novel prize. Second favourite is a biography of Mary Queen of Scots, by John Guy. The fourth book in the running is <i>Corpus</i> , Michael Simmons Roberts' fourth collection of poems. Guy has published many histories, including one of Tudor England. Fletcher has recently graduated from the University of East Anglia's creative writing course.
e is	The Act also applies to firms with more than 300 US shareholders, a situation many firms without US listings could find themselves in. It obliges all firms with US stock market listings to make declarations, which, critics say, will add substantially to the cost of preparing their annual accounts. The US stock market watchdog's chairman has said he is willing to soften tough new US corporate governance rules to ease the burden on foreign firms. The SEC has already delayed implementation of this clause for smaller firms - including US ones - with market capitalisations below \$700m (£374m). Mr Donaldson said the SEC would "consider whether there should be a new approach to the deregistration process" for foreign firms unwilling to meet US requirements. The US regulator said foreign firms may get extra time to comply with a key clause in the Sarbanes-Oxley Act. But European firms with secondary listings in New York have objected - arguing that the compliance costs outweigh the benefits of a dual listing. European firms have protested that US laws introduced after the Enron scandal make Wall Street listings too costly. Compliance costs are already believed to be making firms wary of US listings. Mr Donaldson said that although the Act does not provide exemptions for foreign firms, the Securities and Exchange Commission (SEC) would "continue to be sensitive to the need to accommodate foreign structures and requirements".
jen	She said this would counter "so-called independent" groups like Migration Watch, which she described as an anti-immigration body posing as independent. Migration Watch says it is not against all immigration and the government already publishes accurate figures. She said her proposals mean "we wouldn't have so-called independent experts, like Migration Watch, who come into this debate from an anti-immigration point of view." She went on: "What I would like to see is there being a body which actually looked at the figures, published them, and was independent of government. Barbara Roche said an organisation should monitor and publish figures and be independent of government.
le	The Media Association for America's (MAAA) Free Press File Case reports that their skills are essential for news coverage. The book

Figure 5 5.1.2 Reference English Summary

5.1.3 Adding Manually Written Summaries

In parallel, we manually added human generated English extractive summaries for each Urdu article. These summaries will be used further in rouge score calculation. These scores are compared with the framework-generated summary's score. This evaluation will help us in understanding the efficiency and accuracy of our framework.

The main motivation behind adding this step was to tackle the issue of a low resource language dataset. Since there is no dataset that is specifically available for this study, so we added our own elements and generated this dataset.

D
Manually written summary
Andrea Levy is the favorite to win the White Bread Prize Book of the Year award after winning Novel of the Year with her book Small Island. John Guy's biography of Mary Queen of Scots is another favorite. The judges will decide the overall winner on Tuesday. The five writers in line for the award won their respective categories - debut novel, novel, biography, poetry, and children's book. Other contenders include Susan Fletcher for Eve Green, Michael Simmons Roberts' Corpus, and Geraldine McCaughrean for Not the End of the World. The overall winner will be announced by a panel of judges.
US stock market watchdog chairman William Donaldson has expressed willingness to soften tough corporate governance rules to ease the burden on foreign firms. European firms have protested that US laws introduced after the Enron scandal make Wall Street listings too costly. The Sarbanes-Oxley Act, which comes into force in mid-2005, requires all firms with US stock market listings to make declarations, which critics argue will add substantially to the cost of preparing their annual accounts. Firms that break the law could face huge fines and senior executives risk jail terms of up to 20 years. The SEC is considering a new approach to the deregistration process for foreign firms unwilling to meet US requirements.
Former Home Office minister Barbara Roche has called for an independent body to monitor UK immigration, arguing that it should monitor and publish figures independently of the government. This would counter "so-called independent" groups like Migration Watch, which she describes as anti-

Figure 6 5.1.3 manually written summaries

5.1.4 ROUGE Score Calculation

In order to measure the quantitative value of a summary we use ROUGE to assess this matrix. This will quantitatively assess the similarity between our reference English summary and manually human generated summary.

Here is the breakdown of ROUGE score calculation:

a. ROUGE-1 (Unigram Overlap)

This Rouge matrix is responsible for evaluating the accuracy of word selection. It is done by measuring the intersect of unigrams. Unigrams are individual words between the reference summaries and generated summaries.

b. ROUGE-2 (Bigram Overlap)

Rouge-2 provides an insight into the consistency and continuity of the summary. It is achieved by evaluating the overlap of bigrams, which includes similar sequence of words between the generated and reference summaries.

c. ROUGE-L (Longest Common Subsequence)

To measure the structural similarity between reference and generated summaries we will use Rouge-L. It will calculate the extended common subsequence of words between the both the reference summaries.

d. ROUGE-W (Weighted LCS)

It is a weighted version of ROUGE-L. It considers the length of the longest common subsequence relative to the total number of words in the reference summary.

All of the above Rouge metric contributes to a more detailed evaluation of the quality of summarization.

5.1.4.1 ROUGE Score Calculation Example

Let us discuss a sample example for a sample calculation of ROUGE scores.

Reference summary	Human generated summary
"The government announced new measures to tackle unemployment rates, including job creation programs and incentives for small businesses."	"Government introduces initiatives to combat unemployment, such as job creation schemes and support for small enterprises."

a. ROUGE-1 calculation (Unigram Overlap):

$$\text{ROUGE-1 precision} = \frac{\text{No.of overlapping unigrams in generated summary}}{\text{Total No.of unigrams in generated summary}}$$

$$\text{ROUGE-1 recall} = \frac{\text{No.of overlapping unigrams in generated summary}}{\text{Total No.of unigrams in reference summary}}$$

Overlapping unigrams:

Total 13 overlaps unigrams in reference summary

"government", "introduces", "initiatives", "to", "combat", "unemployment", "such", "as", "job", "creation", "schemes", "and", "for"

Total 18 overlaps unigrams in reference summary:

"the", "government", "announced", "new", "measures", "to", "tackle", "unemployment", "rates", "including", "job", "creation", "programs", "and", "incentives", "for", "small", "businesses"

$$\text{ROUGE-1 precision} = \frac{13}{13} = 1.0$$

$$\text{ROUGE-1 recall} = \frac{13}{18} = 0.7222$$

b. ROUGE-2 calculation (Bigram Overlap):

$$\text{ROUGE-2 precision} = \frac{\text{No.of overlapping bigrams in generated summary}}{\text{Total No.of bigrams in generated summary}}$$

$$\text{ROUGE-2 recall} = \frac{\text{No.of overlapping bigrams in generated summary}}{\text{Total No.of bigrams in reference summary}}$$

Total No. of bigrams in human generated summary: 15

"government introduces", "introduces initiatives", "initiatives to", "to combat", "combat unemployment", "unemployment such", "such as", "as job", "job creation", "creation schemes", "schemes and", "and support", "support for", "for small", "small enterprises"

Total No. of bigrams in reference summary: 17

"the government", "government announced", "announced new", "new measures", "measures to", "to tackle", "tackle unemployment", "unemployment rates", "rates including", "including job", "job creation", "creation programs", "programs and", "and incentives", "incentives for", "for small", "small businesses"

Number of overlapping bigrams: 12

"government introduces", "introduces initiatives", "to combat", "combat unemployment", "unemployment such", "such as", "as job", "job creation", "creation schemes", "schemes and", "and for", "for small"

$$\text{ROUGE-2 precision} = \frac{12}{15} = 0.8$$

$$\text{ROUGE-2 recall} = \frac{12}{17} = 0.7058$$

c. ROUGE-L calculation (Longest common subsequence):

$$\text{ROUGE-L precision} = \frac{\text{Length of LCS}}{\text{Total No.of words in generated summary}}$$

$$\text{ROUGE-L recall} = \frac{\text{Length of LCS}}{\text{Total No.of words in reference summary}}$$

Longest common subsequence:

"government", "introduces", "initiatives", "to", "combat", "unemployment", "such", "as", "job", "creation", "schemes", "and", "for"

$$\text{ROUGE-L precision} = \frac{13}{14} = 0.928$$

$$\text{ROUGE-L recall} = \frac{13}{18} = 0.722$$

d. ROUGE-W calculation (weighted LCS):

$$\text{ROUGE-L precision} = \frac{\text{Length of weighted LCS}}{\text{Total No.of words in generated summary}}$$

$$\text{ROUGE-L recall} = \frac{\text{Length of wieghted LCS}}{\text{Total No.of words in reference summary}}$$

Longest common subsequence:

Same as Rouge-L

$$\text{ROUGE-W precision} = \frac{13}{14} = 0.928$$

$$\text{ROUGE-W recall} = \frac{13}{18} = 0.722$$

The above comprehensive calculations provide precision and recall value for each ROUGE metric. It offers a better insight into the completeness and accuracy of quantitative evaluation of summaries, for our dataset as well as for the framework-generated summaries.

Table 2 5.1.4.1 ROUGE score Calculation

Manual summary rouge score							
Rouge-1		Rouge-2		Rouge-L		Rouge-W	
Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
0.1	0.7222	0.8	0.706	0.928	0.722	0.93	0.722

5.2 Urdu to English Text Translator- Google Translator Library

In the next part of our framework, after setting up the dataset and calculating rouge score for summary evaluation, the next step of our framework is translating Urdu summaries into English using deep translator’s Google translator library. This step is crucial, as it will aid in converting low resource language like URDU to convert into a high resource language. However, the important thing is to check and evaluate the accuracy of this using this part of the code. It is important to keep the original content intact otherwise; it might affect the output of our generated summary.

The google translate library interacts with the google translate API. Which allows it to integrate translation functionality into the framework.

First, we will provide URDU text that we want to translate into English. After that, the library breaks down the text into words or sub words in order to tokenize it. Breaking down the text into smaller linguistic units will prepare the text for translation.

After tokenization, a translation request then initialized by Google Translator library to the API. This translation request includes three essential part. First, one is the input text, which will be the tokenized Urdu text. The second portion contains the source language information that will be URDU in our case. Finally, the third portion contains Target language information, which is English. An http request initiated for this communication between translator library and API.

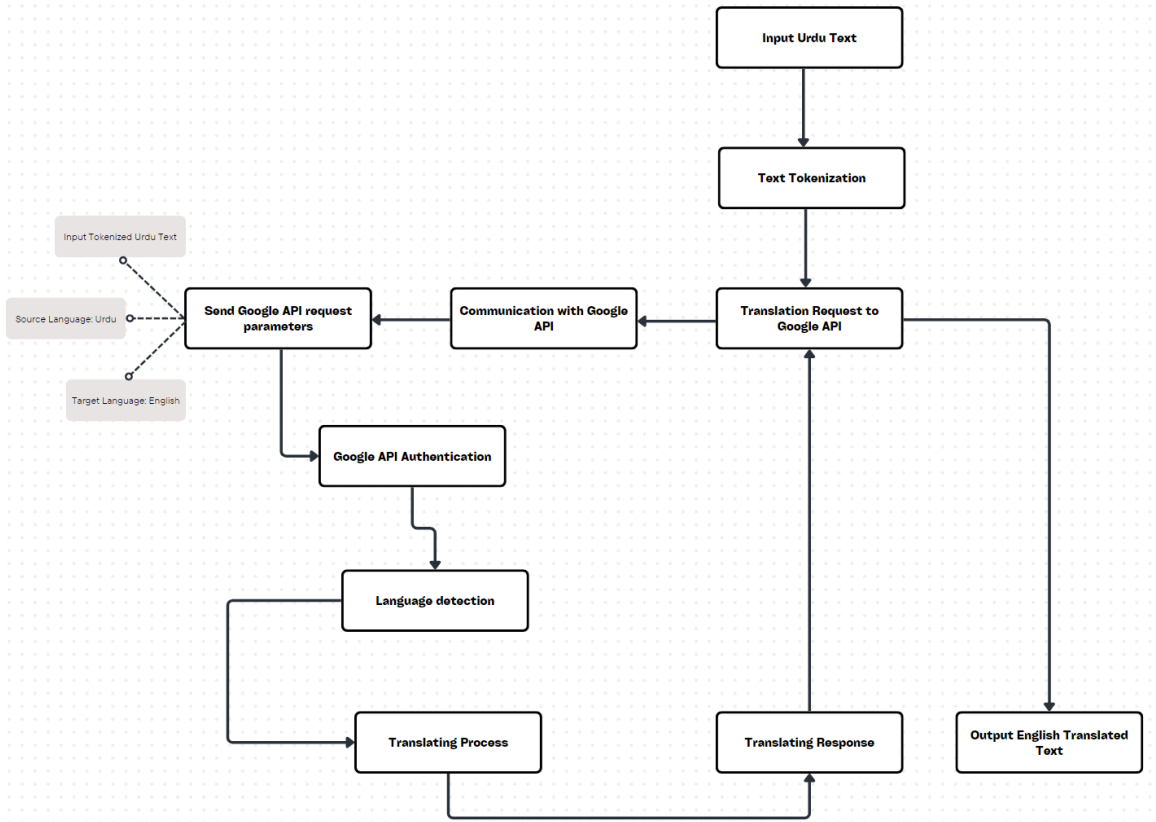


Figure 7 5.2 Translation Process using Google Translator API

After that, the library authenticates the request with the Google translator API using the API key provider. Then, it processes the input text and translates it into desired English output. The API identifies the linguistic elements in the provided Urdu text. It will consider the Urdu language rules along with grammar, syntax and other language dependent factors in order to ensure accuracy and coherency. The translation engine will consider the context of the inputted Urdu text for more accurate results.

The API then sends back the translated output as a response to the initiated translation request. The translator library handles the response and extracts the translated English text from the API, and displays our output as shown in the given example snapshot.

```
from deep_translator import GoogleTranslator

def translate_text(text, source_lang, target_lang):
    translator = GoogleTranslator(source=source_lang, target=target_lang)
    translated_text = translator.translate(text)
    return translated_text

#Urdu text to translate
urdu_text = """
باوجود اس حقیقت کے کہ انگریزی پاکستان میں بہت کم سمجھی اور بولی جاتی ہے
یہ پاکستان کی دفتری زبان ہے اس سے انگریزی کی اہمیت بہت بڑھ گئی ہے۔
انگریزی اس لیے بھی اہم ہے کہ تمام اہم علوم کی کتابیں انگریزی میں لکھی گئی ہیں۔
ہمارے ملک میں انگریزی بولنے والے کو سکاٹر سمجھا جاتا ہے۔
تعلیم کے حصول کے لیے بیرون ملک جانے کے لیے انگریزی کا جاننا بہت ضروری ہے کیونکہ انگریزی بین القوامی زبان ہے۔
تحقیق اور سائنس کے شعبوں میں ترقی کے لیے انگریزی جاننا نہایت ضروری سمجھا جاتا ہے۔
"""

# Translate Urdu to English
translated_text = translate_text(urdu_text, "ur", "en")
print("Translated Text:", translated_text)
```

Translated Text: Despite the fact that English is very little understood and spoken in Pakistan It is the official language of Pakistan, hence the importance of English has increased. English is also important because all important science books are written in English. In our country, an English speaker is considered a scholar. Knowing English is very important to study abroad because English is an international language. Knowing English is considered very important for advancement in the fields of research and science

Figure 8 5.2 Code Snippet of Google Translator

5.3 Text Summarization Using TextRank Algorithm

Going forward with unsupervised text summarization using NLP we will be applying TextRank algorithm. It is an unsupervised NLP algorithm, so it does not require a labeled dataset for training. Instead of that, it will rely on inherent relationships and patterns within the text in order to identify significant information.

This algorithm will leverage graph based ranking methods to extract and summarize keywords. These keywords are extracted based on the semantic similarity and co-occurrences of words and key sentences within the document.

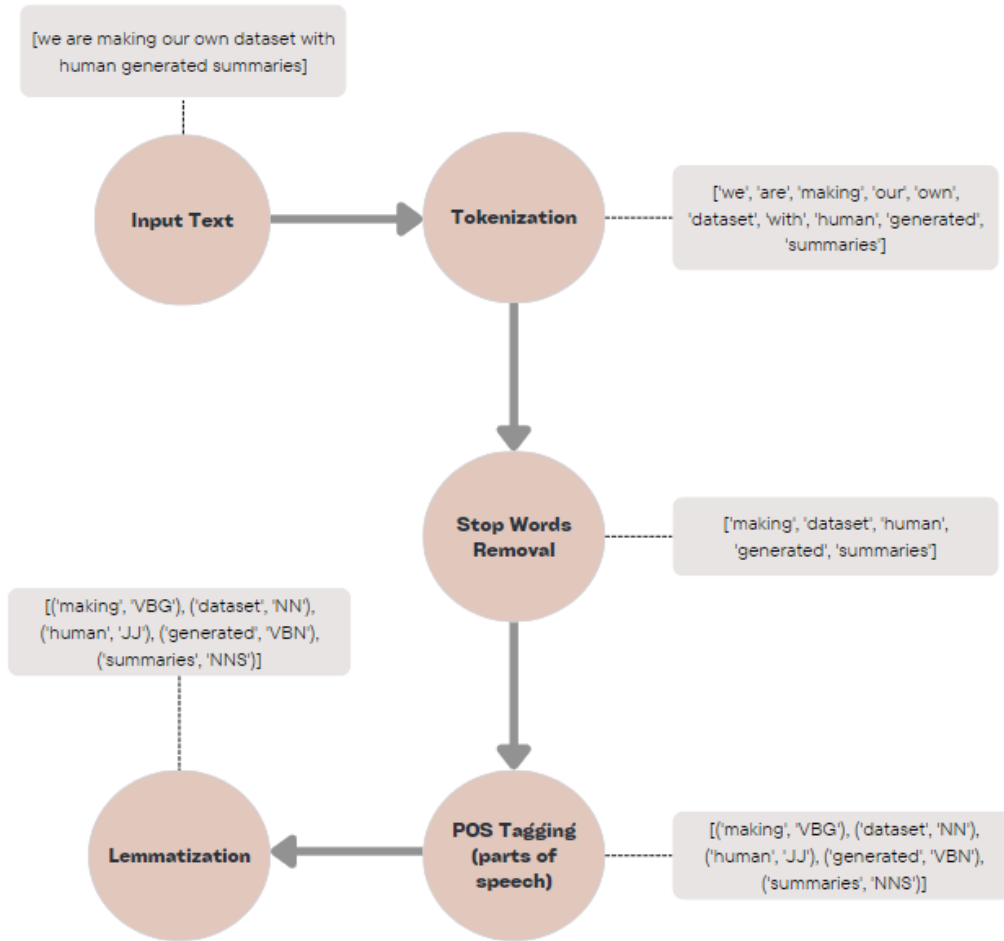


Figure 9 5.3 Text Pre-processing flow chart

Tokenization: the first step will be to tokenize the inputted text into words and sentences. The text will be broken down into smaller linguistic units for further processing. Text representation: the tokenized input text as a graph where each sentence and words are nodes and edges represent the semantic similarity and co-occurrence. Graph representation is constructed by using the same edges.



Figure 10 5.3 TextRank Algorithm working Flow Chart

Edge weighting: after the graph is constructed, the nodes relationships are allotted edge weights. In this case, we are applying cosine similarity as a semantic similarity measure. After that, the Graph Based Ranking will be used to assign importance scored to sentences and words nodes based on their connections to other nodes.

After scoring the algorithm, TextRank will iteratively updates the importance based on the importance scores of neighboring nodes. In order to influence the graph, structure a damping factor is involve in this updating.

This iterative process continues until convergence. At this stage, the importance scores of nodes stabilize.

```
# Preprocess the article
preprocessed_words, original_sentences = preprocess_text(article)

# Calculate TF-IDF scores using CountVectorizer
vectorizer = CountVectorizer()
X = vectorizer.fit_transform([' '.join(words) for words in preprocessed_words])
tfidf_scores = X.toarray()

# Calculate similarity matrix using cosine similarity
similarity_matrix = cosine_similarity(tfidf_scores)

# Apply TextRank algorithm to extract important sentences
def textrank(sentences, similarity_matrix, top_n=6, damping=0.85, max_iter=100):
    scores = np.ones(len(sentences))
    for _ in range(max_iter):
        new_scores = (1 - damping) + damping * (similarity_matrix.T @ (scores / np.sum(similarity_matrix, axis=1)))
        if np.allclose(scores, new_scores, atol=1e-6):
            break
        scores = new_scores
    ranked_indices = np.argsort(-scores)[:top_n]
    return [sentences[i] for i in ranked_indices]
```

Figure 11 5.3 Code Snippet of TextRank

Sentence or Word Importance; higher importance scores of sentences and words are more likely to be included in the summary. For summary extraction, the algorithm will select the sentence with higher importance than the other will. Even if the document is long, this algorithm is robust so it will still identify the key information from any noise and redundancy information in the text.

```
# Apply TextRank to get summary sentences
summary_sentences = textrank(original_sentences, similarity_matrix)

# Print the summary
print("\n".join(summary_sentences))
```

Novelist Andrea Levy is favorite to win the prestigious White Bread Prize Book of the Year a The five writers in line for the award won their respective categories - debut novel, novel, Another favorite is the biography of Mary Queen of Scots, by John Guy. The judges called Guy's My Heart of My Own: The Life of Mary Queen of Scots "an impressive a The book has already won the Orange Prize for Fiction, and is now 5/4 favorite for the £25,0 Other contenders include Susan Fletcher for Eve Green, who won the first novel prize; Fletch [nltk_data] Downloading package punkt to /root/nltk_data...

Figure 12 5.3 Code Snippet of TextRank Algorithm output

5.4 Sentiment Analysis

The fourth component of our framework is adding sentiment analysis. This component does not have a direct impact on the generated summary, instead it is employed to give the readers an insight if the objective aspect of the text. It provides additional context about the summarized text. It enriches the reader's understanding of the subject that is being discussed in our provided text. It informs us whether the content is negative or positive in sense of time, subjective perspective and overall sentiment. It provides a deeper comprehension to our output. Thus, enabling the reader to assess the text's relevance, implications and impact. Also facilitates the interpretation of summarized information.

Since the addition of this part will not directly influence the output of the generated text, but it will enhance the understanding of the text. Sentiment analysis will detect the emotional tone and sentiment polarity expressed within the text. By analyzing the sentiment of the text, the framework can determine the core concerns, emotions and attitude conveyed in the content.

In order to catalogue content into separate sentiment categories, such as +ve, -ve or neutral, natural processing technique is incorporate in sentiment analysis. These sentiments are determined by analyzing sentiment lexicons, contextual cues and linguistic features. In this way, we can enhance the depth of the generated summaries. The framework can provide more holistic representation of the content.

For sentiment analysis algorithm, first we import necessary library. For this algorithm, the code will import NLTK library for sentiment analysis. After that, sentiment analyzer is initialized from NLTK library.

The input text will be tokenized into the sentences. It is one of the crucial steps of sentiment analysis, as it is performed at the sentence level. For each sentence, the algorithm tokenizes the words, and remove stop words. VADER sentiment analyzer function then compound score to represent the overall sentiment of the sentence. Frequency of key phrases identified in the sentence is add to the final count to determine the overall sentiment.

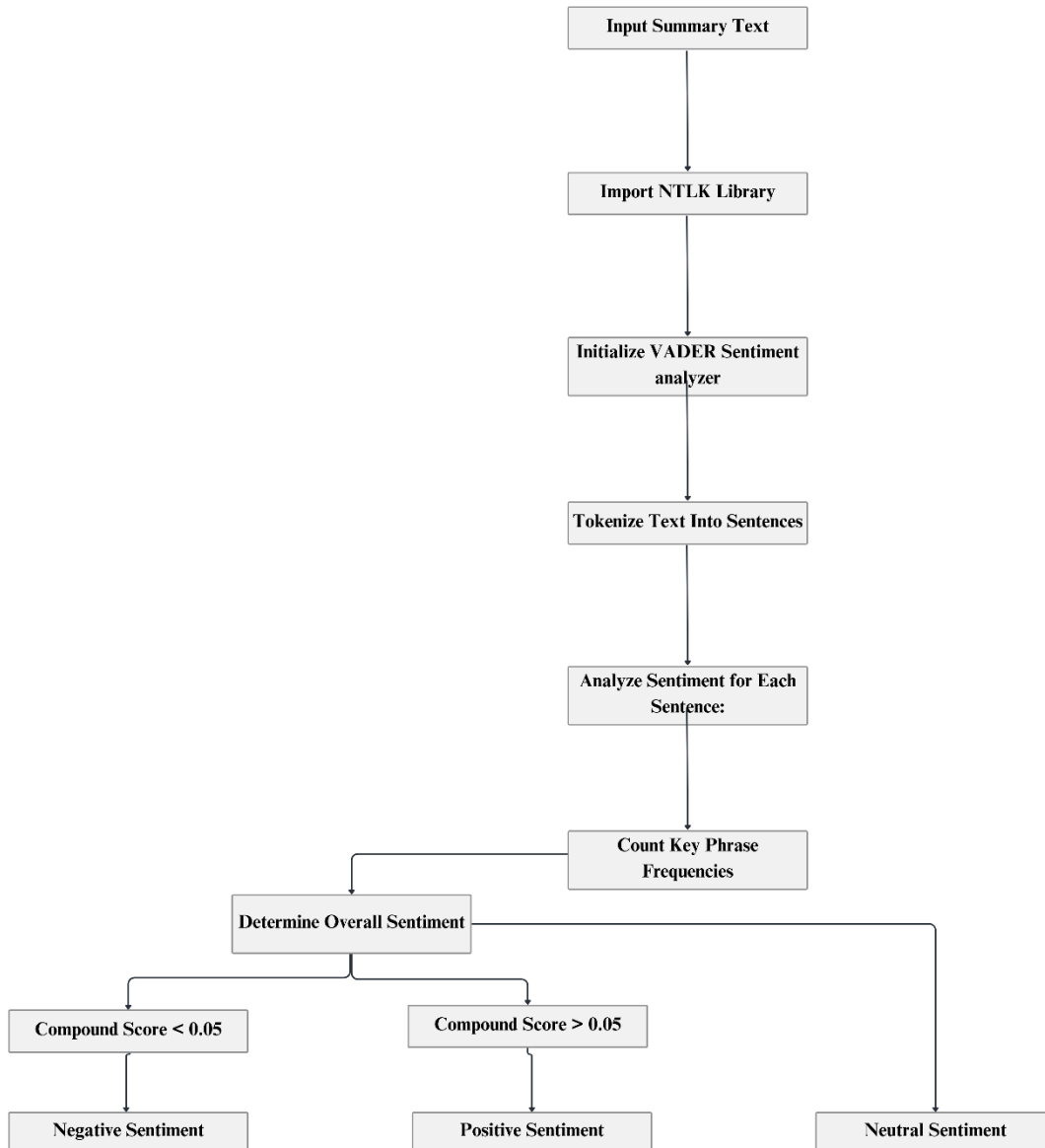


Figure 13 5.4 Sentiment Analysis Flow

If the VADER compound score is greater than 0.05, the output will be positive. If it is less than -0.05, the then it is negative; otherwise, the sentiment is considered neutral. At the end of the output, this algorithm outputs some key words identified in the text that contributes to the sentiment and their respective frequencies.

5.5 ROUGE Score Calculation

To estimate the quality of framework-generated summaries, a set of metrics; ROUGE will be used in this section of the framework. The results of this ROUGE metrics will then be compared to the ROUGE score of reference summaries from over manually curated dataset. ROUGE scores classically comprise of components like recall, precision, F1 scores for n-gram and l-gram matches.

Our reference summary for example usage is as below:

```
# Example usage
reference_summary = """The five writers in line for the award won their respective categories - first novel, novel,
biography, poetry and children's book - on 6 January.The book has already won the Orange Prize for fiction, and
is now 5/4 favourite for the £25,000 Whitbread.Novelist Andrea Levy is favourite to win the main
Whitbread Prize book of the year award, after winning novel of the year with her book Small Island.
The other contenders include Susan Fletcher for Eve Green, which won the first novel prize.Second favourite
is a biography of Mary Queen of Scots, by John Guy.The fourth book in the running is Corpus,
Michael Symmons Roberts' fourth collection of poems.Guy has published many histories, including one of Tudor England.
Fletcher has recently graduated from the University of East Anglia's creative writing course."""
```

Figure 14 .5 Reference summary code snippet

Framework-generated summary for example usage is as below:

```
generated_summary = """Novelist Andrea Levy is favorite to win the prestigious White Bread Prize Book of the
Year award after winning Novel of the Year with her book Small Island.
The five writers in line for the award won their respective categories - debut novel, novel, biography,
poetry and children's book - on January 6.Another favorite is the biography of Mary Queen of Scots,
by John Guy.The judges called Guy's My Heart of My Own: The Life of Mary Queen of Scots "an impressive and
readable piece of scholarship, which cannot fail to leave the reader moved and intrigued by this
most tragic and endearing Queen."The book has already won the Orange Prize for Fiction, and is now 5/4
favorite for the £25,000 Whitbread.Other contenders include Susan Fletcher for Eve Green,
who won the first novel prize; Fletcher recently graduated from the University of East Anglia's
creative writing course."""
```

Figure 15 5.5 Framework-generated summary code snippet

The algorithm will tokenize both the input texts, reference summary and framework generated summary. In order to identify the matching n-grams or l-grams between both the summaries, algorithm compare these tokenized sequences. By aggregating the precision, recall and F1-score across all the l-grams or n-grams in the summaries, it computes the ROUGE scores.

```

def calculate_rouge_scores(reference_summary, generated_summary):
    rouge = Rouge()
    scores = rouge.get_scores(generated_summary, reference_summary)
    return scores

# Example usage

```

Figure 16 5.5 ROUGE score function code snippet

After calculating the ROUGE scores, the output printed for all ROUGE-1, 2 and L scores against each parameter. All of these terminologies have already been explained in the previous chapters.

```

rouge_scores = calculate_rouge_scores(reference_summary, generated_summary)

print("ROUGE Scores:")
print("ROUGE-1 Precision:", rouge_scores[0]['rouge-1']['p'])
print("ROUGE-1 Recall:", rouge_scores[0]['rouge-1']['r'])
print("ROUGE-1 F1-Score:", rouge_scores[0]['rouge-1']['f'])

print("ROUGE-2 Precision:", rouge_scores[0]['rouge-2']['p'])
print("ROUGE-2 Recall:", rouge_scores[0]['rouge-2']['r'])
print("ROUGE-2 F1-Score:", rouge_scores[0]['rouge-2']['f'])

print("ROUGE-L Precision:", rouge_scores[0]['rouge-l']['p'])
print("ROUGE-L Recall:", rouge_scores[0]['rouge-l']['r'])
print("ROUGE-L F1-Score:", rouge_scores[0]['rouge-l']['f'])

```

Figure 17 5.5 Code snippet ROUGE score final step

CHAPTER 6:

6. Results and Analysis

In this chapter, we will evaluate our framework by running it on a sample entry from our manually curated dataset. The framework incorporates various components including, dataset preparation, translating language with limited resources like Urdu to a high resource language English, summarization, sentiment analysis, and evaluation using ROUGE metrics. The analysis of the results we receive from the framework will be mainly focused on comparing the rouge scores between manually generated summary scores and framework generated summary scores. It will provide an insight into the effectiveness of the framework in producing informative and accurate extractive text summarization. For the analysis purposes, we will be moving forward with a chunk of 50 records from our dataset.

Manual Dataset Preparation Results

Our manually curated dataset serves as a benchmark for the evaluation of framework. Since we do not have any prior study results to compare our results with, that is specifically design for Urdu to English cross lingual text summarization. Each textual content is curated carefully in the dataset, along with the human written manual summaries that will capture the key information from the corresponding translated English content.

The analysis of the dataset will eventually lead us to evaluate the quality of our framework generated summaries.

The first column of our dataset contains the original Urdu textual content, while the second column displays the translated English counterparts that is the result of our first part of framework; translation part.

Urdu text
1
2
3
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50

Figure 18 6.0 First column of dataset – Urdu text content

The second column displays the translated English counterparts that is the result of our first part of framework; translation part.


```
[ ] from deep_translator import GoogleTranslator

def translate_text(text, source_lang, target_lang):
    translator = GoogleTranslator(source=source_lang, target=target_lang)
    translated_text = translator.translate(text)
    return translated_text

#Urdu text to translate
urdu_text = """
لیوی نے واٹس بریڈ پرائز کے لیے اشارہ کیا۔
، نگار اینڈریا لیوی اپنی کتاب سمال آئی لینڈ کے ساتھ سال کا بہترین ناول جیتنے کے بعد سال کی ایم واٹس بریڈ پرائز تک آف دی ایئر ایوارڈ جیتنے کے لیے پسندیدہ ہیں۔
ن سوانح عمری ہے، جو جان گائے کی ہے، جوں کا ایک پینل بشمول سر ٹریور میکڈونلڈ، اداکار بیو گرانٹ اور مصنف جوآن بیرس منگل کو مجموعی طور پر فاتح کا فیصلہ کرے گا۔
فین شائع کی ہیں جن میں سے ایک کیوڈر انگلینڈ بھی شامل ہے۔ وہ کلینئر کالج، کیمبرج میں فیلو ہیں اور 2003 میں سینٹ اینڈریوز یونیورسٹی کے اعزازی ریسرچ پروفیسر بنے۔
بجویشن کیا ہے، چل رہی چوتھی کتاب کاریں ہے۔ سائیکل سیمینز رابرٹس کی نظموں کا چوتھا مجموعہ۔ شاعری لکھنے کے ساتھ ساتھ سیمینز رابرٹس دستاویزی فلمیں بھی بناتے ہیں۔
"""

# Translate Urdu to English
translated_text = translate_text(urdu_text, "ur", "en")
print("Translated Text:", translated_text)
```

Figure 19 6.0 Translation code snippet

```
print(translated_text, translated_text)

Translated Text: Levy points for the White Bread Prize.
Novelist Andrea Levy is favorite to win the prestigious White Bread Prize Book of the Year award after winning Novel of the Year with her book Small Island.
The book has already won the Orange Prize for Fiction, and is now 5/4 favorite for the £25,000 Whitbread. Another favorite is the biography of Mary Queen of Scots, by Jo
The five writers in line for the award won their respective categories - debut novel, novel, biography, poetry and children's book - on January 6. Small Island, Levy's f
Other contenders include Susan Fletcher for Eve Green, who won the first novel prize; Fletcher recently graduated from the University of East Anglia's creative writing c
```

Figure 20 6.0 Translation output snippet

Translated English text
The FA is to take action after trouble marred Wednesday's Carling Cup tie between Chelsea and West Disgraced former Chelsea striker Adrian Mutu is to begin talks with Juventus as he looks for a new club Striker Nicolas Anelka reportedly wants to leave Manchester City in search of Champions League football India, which attends the G7 meeting of seven leading industrialised nations on Friday, is unlikely to be Indonesia's government has confirmed it is considering raising fuel prices by as much as 30%. Million India has raised the limit for foreign direct investment in telecoms companies from 49% to 74%. Comm The French economy picked up speed at the end of 2004, official figures show - but still looks set to ha The gap between US exports and imports hit an all-time high of \$671.7bn (£484bn) in 2004, latest figur The creator of Buffy the Vampire Slayer is to take on a new female superhero after signing up to write Lord of the Rings director Peter Jackson has said that it will be up to four years before he starts work on British film director Sir Alan Parker has been made an officer in the Order of Arts and Letters, one of F The sixth and final Star Wars movie may not be suitable for young children, film-maker George Lucas The low-budget horror film Boogeyman has knocked Robert de Niro thriller Hide and Seek from the top

Figure 21 6.0 Second column of dataset – English translated text

Next column includes the reference English summary. The manually written summaries is in the next column, as seen in the attached below snapshot. The reference summary serves as a reference point for comparison.

C	
1	Reference English Summary
2	
3	
28	West Ham boss Alan Pardew said: "It's a shame because I thought there wa
29	Disgraced former Chelsea striker Adrian Mutu is to begin talks with Juventu
30	Playing for eighth place is good but I miss the Champions League. Anelka, ;
31	At a conference on developing enterprise hosted by UK finance minister Gc
32	Indonesia's government has confirmed it is considering raising fuel prices
33	We need at least \$20bn (£10.6bn) in investment and part of this has to come
34	Despite the apparent shortfall in annual economic growth, the good quarter!
35	The Commerce Department said the trade deficit for all of last year was 24.4

Figure 22 6.0 Third column of dataset – Reference English summary

A	
1	Manually written summary
2	
3	
4	The favourite character to win the Whote bread Prize is considered to be Andres Levy. Especially after she wrote the her new book. But regardless of what people think, the final reuslt will be drawn by the judges.
5	There is apossibility of showin some lineancy towards the governance rules for foreign firms. The willingness is epxressd by the US stock maerket chairman.
6	Due to overcrowding and culture former home minster barbara has shown willingness in forming a body to regulate US immigration laws.

Figure 23 6.0 manually written summaries column

The next few columns in the results screenshot depict the calculation of ROUGE scores. These scores calculated between manually written summaries against the reference summaries.

```

from rouge import Rouge

def calculate_rouge_scores(reference_summary, generated_summary):
    rouge = Rouge()
    scores = rouge.get_scores(generated_summary, reference_summary)
    return scores

# Example usage
reference_summary = """Buhecha was previously a legitimate distributor of Bollywood films, but was suspended and sued by his employers for dealing in illegal copies of Bollywood classic Mohabbatein. Buhecha, who made £26,000 per month from his illegal trade, was called "one of the biggest Bollywood pirates in the UK" by the sentencing judge. A major distributor of pirated DVDs of Bollywood films has been sent to prison for three years. The judge in the case, which lasted seven days, said that "a heavy penalty was called for because of the enormous damage Buhecha caused to legitimate business". An operation was launched against Buhecha in 2002 after complaints were received about his activities. Jayanti Amarishi Buhecha from Cambridge was found guilty of two trademark offences last month, and sentenced at Harrow Crown Court, London, on Tuesday. Legitimate Bollywood film distributors have hailed the conviction as "a major boost".
"""

manual_summary = """Jayanti Amarishi Buhecha, a distributor of pirated Bollywood DVDs, has been sentenced to three years in prison. Buhecha, who earned £26,000 per month from his illegal trade, was deemed one of the largest Bollywood pirates in the UK. The British Phonographic Industry (BPI) worked on the case for two years, and a heavy penalty was imposed due to the damage caused to legitimate businesses. Buhecha was previously a legitimate distributor of Bollywood films but was suspended and sued by his employers for dealing in illegal copies of Mohabbatein. The BPI welcomed the prison sentence but warned that there are many other active counterfeiters of Bollywood films. The BPI's anti-piracy director, David Martin, warned that the problem will not disappear with Buhecha, and it is crucial to continue efforts in this field.
"""

```

Figure 24 6.0 Manual summaries rouge score calculation code snippet-1

```
rouge_scores = calculate_rouge_scores(reference_summary, manual_summary)

print("ROUGE Scores:")
print("ROUGE-1 Precision:", rouge_scores[0]['rouge-1']['p'])
print("ROUGE-1 Recall:", rouge_scores[0]['rouge-1']['r'])
print("ROUGE-1 F1-Score:", rouge_scores[0]['rouge-1']['f'])
print("\r\n")
print("ROUGE-2 Precision:", rouge_scores[0]['rouge-2']['p'])
print("ROUGE-2 Recall:", rouge_scores[0]['rouge-2']['r'])
print("ROUGE-2 F1-Score:", rouge_scores[0]['rouge-2']['f'])
print("\r\n")
print("ROUGE-L Precision:", rouge_scores[0]['rouge-l']['p'])
print("ROUGE-L Recall:", rouge_scores[0]['rouge-l']['r'])
print("ROUGE-L F1-Score:", rouge_scores[0]['rouge-l']['f'])
```

Figure 25 6.0 Manual summaries rouge score calculation code snippet-2

```
ROUGE Scores:
ROUGE-1 Precision: 0.5376344086021505
ROUGE-1 Recall: 0.47619047619047616
ROUGE-1 F1-Score: 0.5050505000688705

ROUGE-2 Precision: 0.3228346456692913
ROUGE-2 Recall: 0.2949640287769784
ROUGE-2 F1-Score: 0.30827067170190514

ROUGE-L Precision: 0.5161290322580645
ROUGE-L Recall: 0.45714285714285713
ROUGE-L F1-Score: 0.48484847986685037
```

Figure 26 6.0 Manual summaries rouge score calculation output snippet

This step is crucial, with the help of it we will be able to measure the similarity between these two summaries in terms of recall, precision and F1-score. The below mentioned table snippet shows the calculated rouge score values of manual written summaries.

Manual summary rouge score								
Rouge-1			Rouge-2			Rouge-L		
Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1score
0.6350	0.5100	0.5660	0.3750	0.2720	0.3150	0.5940	0.4780	0.5300
0.6880	0.4330	0.5320	0.5270	0.2660	0.3530	0.6660	0.4190	0.3530
0.4920	0.4070	0.4460	0.2027	0.1530	0.1744	0.4920	0.4078	0.4460
0.753	0.5680	0.6480	0.5240	0.4029	0.4550	0.7402	0.5580	0.6360
0.4070	0.3020	0.3460	0.1527	0.1208	0.1340	0.3510	0.2602	0.2990
0.7846	0.6456	0.7083	0.6835	0.5094	0.5838	0.7538	0.6203	0.6806
0.5244	0.4574	0.4886	0.3445	0.3228	0.3333	0.5244	0.4574	0.4886
0.6500	0.6566	0.6533	0.4597	0.4286	0.4436	0.6300	0.6364	0.6332
0.6341	0.4727	0.5417	0.4375	0.3182	0.3684	0.6341	0.4727	0.5417
0.5843	0.5361	0.5591	0.3583	0.3209	0.3386	0.5730	0.5258	0.5484
0.3600	0.3462	0.3529	0.1930	0.1719	0.1818	0.3000	0.2885	0.2941
0.8222	0.5362	0.6491	0.6275	0.3810	0.4741	0.8000	0.5217	0.6316
0.4677	0.5577	0.5088	0.2169	0.2769	0.2432	0.4516	0.5385	0.4912
0.4563	0.4352	0.4455	0.2824	0.2517	0.2662	0.4563	0.4352	0.4455
0.6264	0.5135	0.5644	0.3790	0.3264	0.3507	0.5934	0.4865	0.5347
0.3134	0.3818	0.3443	0.1304	0.1385	0.1343	0.3134	0.3818	0.3443
0.4250	0.5231	0.4690	0.2661	0.3258	0.2929	0.4125	0.5077	0.4552
0.4565	0.5122	0.4828	0.3208	0.3617	0.3400	0.4348	0.4878	0.4598
0.4138	0.4045	0.4091	0.2177	0.2126	0.2151	0.3908	0.3820	0.3864
0.6140	0.5072	0.5556	0.4028	0.3222	0.3580	0.5965	0.4928	0.5397
0.6271	0.5211	0.5692	0.4783	0.3793	0.4231	0.6271	0.5211	0.5692
0.7170	0.6786	0.6972	0.5077	0.5000	0.5038	0.6981	0.6607	0.6789
0.4769	0.4397	0.4576	0.3115	0.2767	0.2931	0.4769	0.4397	0.4576
0.6190	0.5306	0.5714	0.3878	0.2969	0.3363	0.6190	0.5306	0.5714
0.3854	0.3491	0.3663	0.2000	0.1778	0.1882	0.3646	0.3302	0.3465
0.4884	0.4286	0.4565	0.2128	0.1818	0.1961	0.4419	0.3878	0.4130
0.5111	0.5349	0.5227	0.2101	0.2212	0.2155	0.5111	0.5349	0.5227
0.6699	0.5948	0.6301	0.4662	0.4052	0.4336	0.6408	0.5690	0.6027
0.6129	0.6706	0.6404	0.4839	0.4839	0.4839	0.5914	0.6471	0.6180
0.5205	0.5278	0.5241	0.2660	0.2778	0.2717	0.4932	0.5000	0.4966
0.5510	0.3913	0.4576	0.2969	0.2317	0.2603	0.5306	0.3768	0.4407
0.5888	0.5625	0.5753	0.3623	0.3185	0.3390	0.5327	0.5089	0.5205
0.4487	0.3723	0.4070	0.2500	0.1866	0.2137	0.4231	0.3511	0.3837
0.5057	0.5238	0.5146	0.3455	0.3220	0.3333	0.4943	0.5119	0.5029
0.5610	0.6866	0.6174	0.3810	0.4651	0.4188	0.5122	0.6269	0.5638

Figure 27 6.0 ROUGE scores of manually written summaries

The ROUGE evaluation provides insights into the association between these two summaries. These scores will be further use for comparing it to assess the effectiveness of framework-generated summaries.

Moving forward, this set of results screenshots represents the framework-generated summaries in the subsequent column. These summaries are produced using our cross lingual text summarizer framework, incorporating a translator and unsupervised NLP technique: TextRank.

```

import numpy as np
from nltk.corpus import stopwords
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.stem import PorterStemmer
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.feature_extraction.text import CountVectorizer

nltk.download('punkt')
nltk.download('stopwords')

# Sample article text
article = """A major distributor of pirated DVDs of Bollywood films has been sent to prison for three years. Jayanti Amarishi Buhecha from Cambridge was found guilty of two trademark offences last month, and sentenced at Harrow Crown Court, London, on Tuesday. Buhecha, who made £26,000 per month from his illegal trade, was called "one of the biggest Bollywood pirates in the UK" by the sentencing judge. The British Phonographic Industry (BPI) worked for two years on the case. An operation was launched against Buhecha in 2002 after complaints were received about his activities. The judge in the case, which lasted seven days, said that "a heavy penalty was called for because of the enormous damage Buhecha caused to legitimate business". Fake DVDs were manufactured in Pakistan and Malaysia and sold on wholesale to shops by Buhecha, who traded in counterfeit DVDs in 2002 and 2003. In December 2002, he was stopped in his car by trading standards officers, who uncovered 1,000 pirated DVDs and faked inlay cards printed with registered trademarks. Despite being arrested and bailed, Buhecha was caught a second time at the end of 2003. His home and a lock-up in Cambridge were found to contain 18,000 counterfeit DVDs and further faked inlay cards. Buhecha was previously a legitimate distributor of Bollywood films, but was suspended and sued by his employers for dealing in illegal copies of Bollywood classic Mohabbatein. Legitimate Bollywood film distributors have hailed the conviction as "a major boost". Bollywood music and film suffers piracy at the rate of 40%, which is more than that suffered by mainstream productions. The BPI welcomed the news of the prison sentence, but warned there are plenty of other active counterfeiters of Bollywood films. The organisation's anti-piracy director David Martin said: "The problem simply will not disappear with Buhecha. Others and more will take his place, so it's vital that keep up our efforts in this field."
"""

```

Figure 28 6.0 TextRank algorithm code-1

```

# Preprocessing
def preprocess_text(text):
    sentences = sent_tokenize(text)
    words = [word_tokenize(sentence) for sentence in sentences]
    words = [[word.lower() for word in sentence if word.isalnum()] for sentence in words]
    stop_words = set(stopwords.words('english'))
    words = [[word for word in sentence if word not in stop_words] for sentence in words]
    ps = PorterStemmer()
    words = [[ps.stem(word) for word in sentence] for sentence in words]
    return words, sentences

# Preprocess the article
preprocessed_words, original_sentences = preprocess_text(article)

# Calculate TF-IDF scores using CountVectorizer
vectorizer = CountVectorizer()
X = vectorizer.fit_transform([' '.join(words) for words in preprocessed_words])
tfidf_scores = X.toarray()

# Calculate similarity matrix using cosine similarity
similarity_matrix = cosine_similarity(tfidf_scores)

# Apply TextRank algorithm to extract important sentences
def textrank(sentences, similarity_matrix, top_n=5, damping=0.85, max_iter=100):
    scores = np.ones(len(sentences))

```

Figure 29 6.0 TextRank algorithm code-2

```

ranked_indices = np.argsort(-scores)[:top_n]
return [sentences[i] for i in ranked_indices]

# Apply TextRank to get summary sentences
summary_sentences = textrank(original_sentences, similarity_matrix)

# Print the summary
print("\n".join(summary_sentences))

```

A major distributor of pirated DVDs of Bollywood films has been sent to prison for three years. Buhecha was previously a legitimate distributor of Bollywood films, but was suspended and sued by his employers for dealing in illegal copies of Bollywood classic Mohabbatein. Buhecha, who made £26,000 per month from his illegal trade, was called "one of the biggest Bollywood pirates in the UK" by the sentencing judge. Fake DVDs were manufactured in Pakistan and Malaysia and sold on wholesale to shops by Buhecha, who traded in counterfeit DVDs in 2002 and 2003. The BPI welcomed the news of the prison sentence, but warned there are plenty of other active counterfeiters of Bollywood films.

Figure 30 6.0 Framework-generated text output

A major distributor of pirated DVDs of Bollywood films has been sent to prison for three years. Buhecha was previously a legitimate distributor of Bollywood films, but was suspended and sued by his employers for dealing in illegal copies of Bollywood classic Mohabbat. Buhecha, who made £26,000 per month from his illegal trade, was called "one of the biggest Bollywood pirates in the UK" by the sentencing judge. Fake DVDs were manufactured in Pakistan and Malaysia and sold on wholesale to shops by Buhecha, who traded in counterfeit DVDs in 2002 and 2003. The BPI welcomed the news of the prison sentence, but warned there are plenty of other active counterfeiters of Bollywood films.

Figure 31 Figure 14-6.0 Framework-generated text column

The next set of screenshots contains Sentimental analysis part of the framework. As mentioned earlier, this part of the framework will not have a direct impact on the output of the generated summary. However, this part of the framework will provide a deeper insight into the overall context of the generated summary. It tells the reader about the essence of the text, either its positive, negative or neutral content. Along with that, it provides a little more detail into those keywords that sums up the sentiment analysis result. Our code will be implementing VADER sentiment analyser.

```

"""
Analyzes the sentiment of the given text using VADER sentiment analysis tool.

Args:
- text (str): The text to analyze.

Returns:
- sentiment (str): The sentiment of the text ('positive', 'negative', or 'neutral').
- explanation (str): A brief explanation of the sentiment based on key phrases.
"""
# Initialize the VADER sentiment analyzer
sid = SentimentIntensityAnalyzer()

# Tokenize the text into sentences
sentences = sent_tokenize(text)

# Analyze sentiment for each sentence and extract key phrases
key_phrases = []
for sentence in sentences:
    # Tokenize words in the sentence and remove stopwords
    words = word_tokenize(sentence)
    words = [word.lower() for word in words if word.isalpha() and word.lower() not in stopwords.words('

```

Figure 32 6.0 Code snippet for sentiment analysis

```

# Example text to analyze
text = """Novelist Andrea Levy is favorite to win the prestigious White Bread Prize Book of the Year aw
The five writers in line for the award won their respective categories - debut novel, novel, biography, poe
Another favorite is the biography of Mary Queen of Scots, by John Guy.
The judges called Guy's My Heart of My Own: The Life of Mary Queen of Scots "an impressive and readable pie
The book has already won the Orange Prize for Fiction, and is now 5/4 favorite for the £25,000 Whitbread.
Other contenders include Susan Fletcher for Eve Green, who won the first novel prize; Fletcher recently gra

# Analyzing the sentiment of the text
sentiment, explanation = analyze_sentiment(text)

# Printing the sentiment and explanation
print("Sentiment:", sentiment)
print("Explanation:", explanation)

```


 Sentiment: positive
Explanation: The sentiment is positive due to the presence of key phrases: book (4 times), novel (4 times),

Figure 33 6.0 Output for sentiment analysis code

The final set of screenshot results shows the ROUGE scores calculations. These calculations are similar to the manual summaries. The scores represent the quantification of similarity between the reference summary and framework-generated summaries. The next step shows the comparison between the ROUGE scores of manually written summaries and framework written summaries.

```

from rouge import Rouge

def calculate_rouge_scores(reference_summary, generated_summary):
    rouge = Rouge()
    scores = rouge.get_scores(generated_summary, reference_summary)
    return scores

# Example usage
reference_summary = """Buhecha was previously a legitimate distributor of Bollywood films, but was suspended and sued by his employers for dealing in illegal copies of Bollywood classic Mohabbatein.Buhecha, who made £26,000 per month from his illegal trade, was called "one of the biggest Bollywood pirates in the UK" by the sentencing judge.A major distributor of pirated DVDs of Bollywood films has been sent to prison for three years. The judge in the case, which lasted seven days, said that "a heavy penalty was called for because of the enormous damage Buhecha caused to legitimate business". An operation was launched against Buhecha in 2002 after complaints were received about his activities.Jayanti Amarishi Buhecha from Cambridge was found guilty of two trademark offences last month, and sentenced at Harrow Crown Court, London, on Tuesday.Legitimate Bollywood film distributors have hailed the conviction as "a major boost".
"""

framework_generated_summary = """A major distributor of pirated DVDs of Bollywood films has been sent to prison for three years. Buhecha was previously a legitimate distributor of Bollywood films, but was suspended and sued by his employers for dealing in illegal copies of Bollywood classic Mohabbatein. Buhecha, who made £26,000 per month from his illegal trade, was called "one of the biggest Bollywood pirates in the UK" by the sentencing judge. Fake DVDs were manufactured in Pakistan and Malaysia and sold on wholesale to shops by Buhecha, who traded in counterfeit DVDs in 2002 and 2003. The BPI welcomed the news of the prison sentence, but warned there are plenty of other active counterfeiters of Bollywood films.
"""

```

Figure 34 6.0 Rouge score calculation code snippet-1

```

rouge_scores = calculate_rouge_scores(reference_summary, framework_generated_summary)

print("ROUGE Scores:")
print("ROUGE-1 Precision:", rouge_scores[0]['rouge-1']['p'])
print("ROUGE-1 Recall:", rouge_scores[0]['rouge-1']['r'])
print("ROUGE-1 F1-Score:", rouge_scores[0]['rouge-1']['f'])
print("\r\n")
print("ROUGE-2 Precision:", rouge_scores[0]['rouge-2']['p'])
print("ROUGE-2 Recall:", rouge_scores[0]['rouge-2']['r'])
print("ROUGE-2 F1-Score:", rouge_scores[0]['rouge-2']['f'])
print("\r\n")
print("ROUGE-L Precision:", rouge_scores[0]['rouge-l']['p'])
print("ROUGE-L Recall:", rouge_scores[0]['rouge-l']['r'])
print("ROUGE-L F1-Score:", rouge_scores[0]['rouge-l']['f'])

```

Figure 35 6.0 Rouge score calculation code snippet-2

```

ROUGE Scores:
ROUGE-1 Precision: 0.7272727272727273
ROUGE-1 Recall: 0.5333333333333333
ROUGE-1 F1-Score: 0.6153846105029586

ROUGE-2 Precision: 0.5981308411214953
ROUGE-2 Recall: 0.460431654676259
ROUGE-2 F1-Score: 0.5203251983366384

ROUGE-L Precision: 0.7142857142857143
ROUGE-L Recall: 0.5238095238095238
ROUGE-L F1-Score: 0.6043955995139477

```

Figure 36 6.0 Rouge score output for framework-generated summary

framework generated summary rouge score								
Rouge-1			Rouge-2			Rouge-L		
Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1score
0.6470	0.7390	0.6900	0.5000	0.5450	0.5210	0.6380	0.7280	0.6800
1.0000	0.6710	0.8033	0.9630	0.6050	0.7430	1.0000	0.6710	0.8030
0.9866	0.9730	0.9801	0.9690	0.9693	0.9693	0.9866	0.9730	0.9801
0.9600	0.9410	0.9504	0.9090	0.8950	0.9020	0.9600	0.9411	0.9504
0.8580	0.9170	0.8870	0.7840	0.8790	0.8290	0.8460	0.9040	0.8740
0.7684	0.9241	0.8391	0.7176	0.8868	0.7932	0.7684	0.9241	0.8391
0.8354	0.7021	0.7630	0.8077	0.6614	0.7273	0.8228	0.6915	0.7514
1.0000	1.0000	1.0000	0.9850	0.9850	0.9850	1.0000	1.0000	1.0000
0.6667	0.9818	0.7941	0.6500	0.9848	0.7831	0.6667	0.9818	0.7941
1.0000	0.7938	0.8851	0.9906	0.7836	0.8750	1.0000	0.7938	0.8851
0.9535	0.7885	0.8632	0.9000	0.7031	0.7895	0.9535	0.7885	0.8632
0.6111	0.4783	0.5366	0.5455	0.4286	0.4800	0.6111	0.4783	0.5366
0.9714	0.6538	0.7816	0.9318	0.6308	0.7523	0.9714	0.6538	0.7816
0.7900	0.7315	0.7596	0.7518	0.7007	0.7254	0.7900	0.7315	0.7596
0.7976	0.6036	0.6872	0.7182	0.5486	0.6220	0.7976	0.6036	0.6872
0.8209	1.0000	0.9016	0.7683	0.9692	0.8571	0.8209	1.0000	0.9016
0.8000	0.9846	0.8828	0.7500	0.9438	0.8358	0.8000	0.9846	0.8828
0.5200	0.9512	0.6724	0.4490	0.9362	0.6069	0.5200	0.9512	0.6724
0.9831	0.6517	0.7838	0.9506	0.6063	0.7404	0.9831	0.6517	0.7838
0.6420	0.7536	0.6933	0.5981	0.7111	0.6497	0.6420	0.7536	0.6933
0.7969	0.7183	0.7556	0.7284	0.6782	0.7024	0.7969	0.7183	0.7556
0.8485	1.0000	0.9180	0.7875	0.9545	0.8630	0.8485	1.0000	0.9180
0.9815	0.7518	0.8514	0.9412	0.6990	0.8022	0.9815	0.7518	0.8514
1.0000	1.0000	1.0000	0.9531	0.9531	0.9531	1.0000	1.0000	1.0000
0.7013	0.5094	0.5902	0.5699	0.3926	0.4649	0.6883	0.5000	0.5792
1.0000	1.0000	1.0000	0.9455	0.9455	0.9455	1.0000	1.0000	1.0000
0.7500	0.6977	0.7229	0.6449	0.6106	0.6273	0.7375	0.6860	0.7108
0.9063	0.7500	0.8208	0.8167	0.6405	0.7179	0.9063	0.7500	0.8208
0.7595	0.7059	0.7317	0.7364	0.6532	0.6923	0.7595	0.7059	0.7317
0.7368	0.7778	0.7568	0.6337	0.7111	0.6702	0.7368	0.7778	0.7568
0.7887	0.8116	0.8000	0.7079	0.7683	0.7368	0.7887	0.8116	0.8000
0.8081	0.7143	0.7583	0.7518	0.6561	0.7007	0.7980	0.7054	0.7488
0.8837	0.8085	0.8444	0.8609	0.7388	0.7952	0.8837	0.8085	0.8444
1.0000	0.9881	0.9940	0.9744	0.9661	0.9702	1.0000	0.9881	0.9940
0.8442	0.9701	0.9028	0.8265	0.9419	0.8804	0.8442	0.9701	0.9028
1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.7872	1.0000	0.8810	0.6984	0.9565	0.8073	0.7872	1.0000	0.8810
0.7821	1.0000	0.8777	0.7374	0.9605	0.8343	0.7821	1.0000	0.8777
0.6341	0.5200	0.5714	0.5088	0.4085	0.4531	0.6220	0.5100	0.5604
0.8375	0.7976	0.8171	0.7885	0.7455	0.7664	0.8375	0.7976	0.8171
1.0000	0.7865	0.8805	0.9677	0.7317	0.8333	1.0000	0.7865	0.8805
0.6883	0.6974	0.6928	0.6000	0.5745	0.5870	0.6883	0.6974	0.6928

Figure 37 6.0 Calculated ROUGE scores for framework-generated summaries

The major focus of the result analysis is a comparison between the ROUGE scores of framework-generated summaries against the scores of manually written summaries. In this way, we can have a better look at any similarities and discrepancies in the ROUGE score.

As we can see the data is providing an insight into the scores, framework generated ROUGE score is closer to our original reference summaries, which makes the our framework efficient enough to provide summaries that has relevance, accuracy and information as in the original text.

Manual - Rouge 1 Precision	Manual - Rouge 1 Recall	Manual - Rouge 1 F1 score	Manual - Rouge 2 Precision	Manual - Rouge 2 Recall	Manual - Rouge 2 F1 score	Manual - Rouge L Precision	Manual - Rouge L Recall	Manual - Rouge L F1score
0.54	0.50	0.52	0.34	0.30	0.32	0.52	0.48	0.49

Manual Summary								
Rouge-1			Rouge-2			Rouge-L		
Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
0.54	0.50	0.52	0.34	0.30	0.32	0.52	0.48	0.49

Figure 38 6.0 Average of each rouge score for manual summaries

The mentioned above figure shows the average for first 50 records of Rouge scores of manual and framework generated summaries. The data is organized into columns; each represents the Rouge score’s corresponding average value. By visualizing these values, we can see the effectiveness of the summarization process of our framework generated summaries.

Framework - Rouge 1 Precision	Framework - Rouge 1 Recall	Framework - Rouge 1 F1 score	Framework - Rouge 2 Precision	Framework - Rouge 2 Recall	Framework - Rouge 2 F1 score	Framework - Rouge L Precision	Framework - Rouge L Recall	Framework - Rouge L F1score
0.84	0.79	0.80	0.78	0.74	0.75	0.83	0.79	0.80

Framework Generated Summary								
Rouge-1			Rouge-2			Rouge-L		
Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
0.84	0.79	0.80	0.78	0.74	0.75	0.83	0.79	0.80

Figure 39 6.0 Average of each rouge score for generated summaries

A higher score of framework generated Rouge score indicates a better alignment and similarity between the generated summary and our original content.

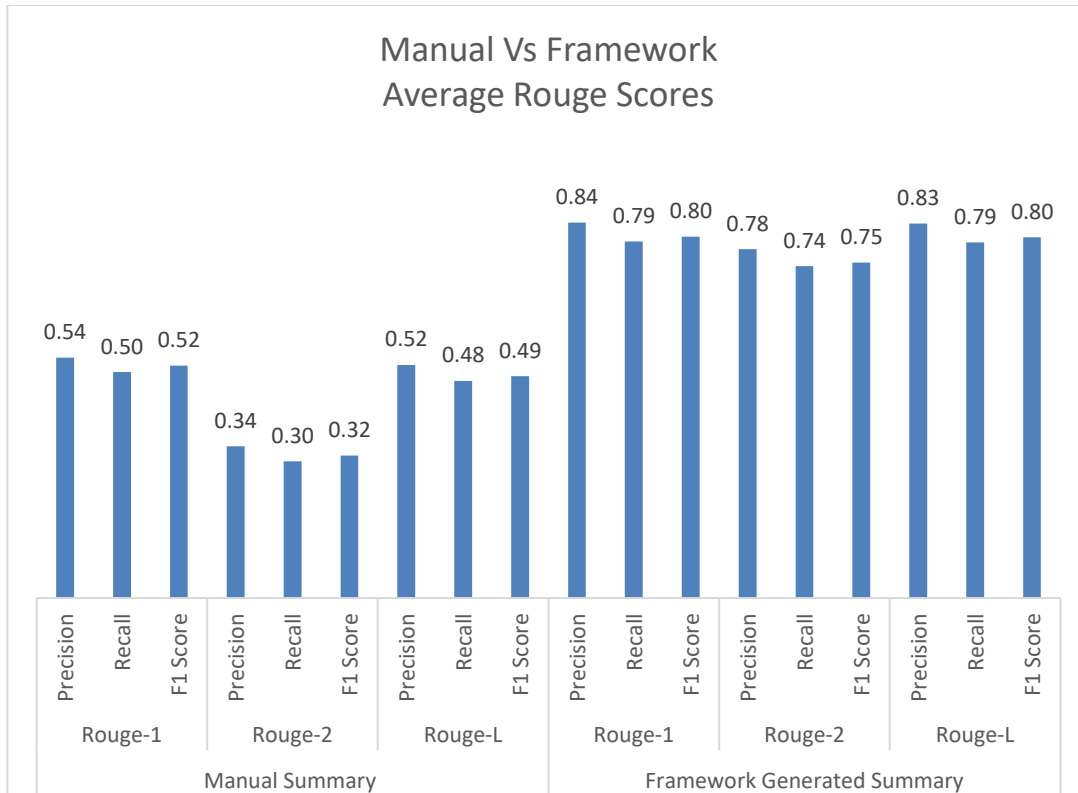


Figure 40 6.0 Rouge scores comparison bar graph

The bar graph above illustrates the comparison between the average Rouge score of first 50 entries of our dataset for both manual written summaries and framework-generated summaries.

The x-axis of the graph shows each ROUGE score along with precision, recall and F1. The graphical visualization shows that framework-generated scores are consistently higher as compared to that of manual summary score.

The next *Figure 24-6.0* line chart compares the precision trend of ROUGE-1 scores, between average values of manual and framework-generated summaries.

The trend shows a clear higher consistency of framework generated precision values. The Y-axis represents the precision scores. The chart includes two trend lines, red one for framework-generated summary scores, and blue for manual summary rouge scores.

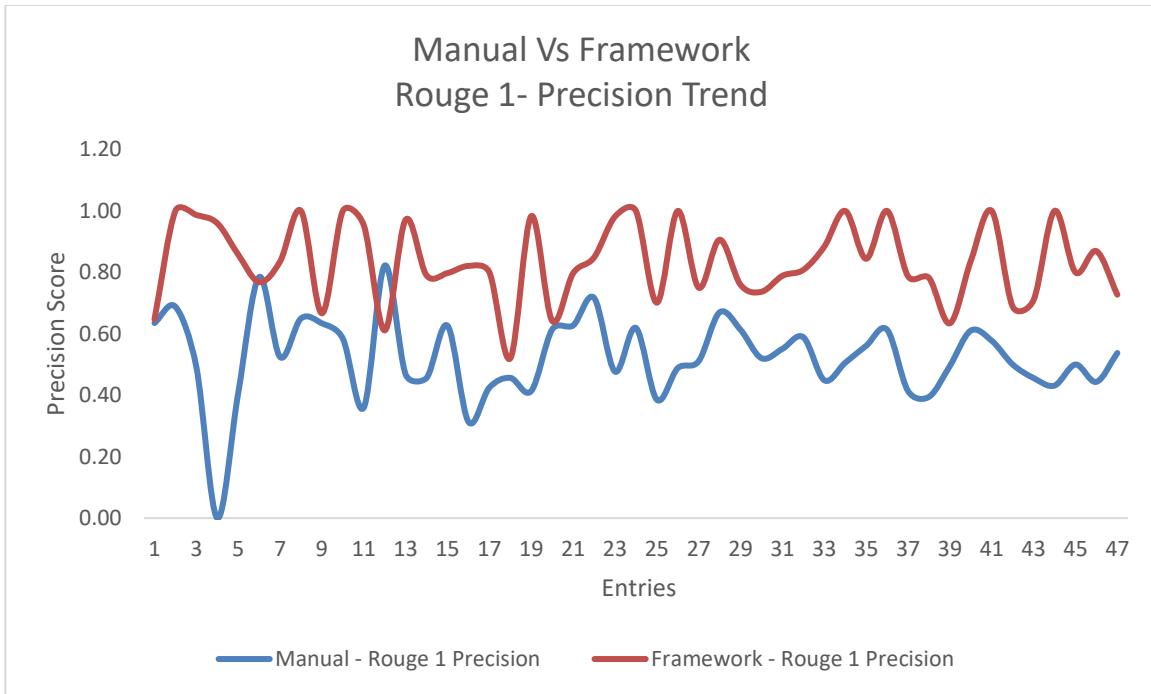


Figure 41 6.0 Comparison line graph for ROUGE-1 precision trend

The below *Figure 25-6.0* line chart compares the F1 scores trend of ROUGE-1, between average values of manual and framework-generated summaries.

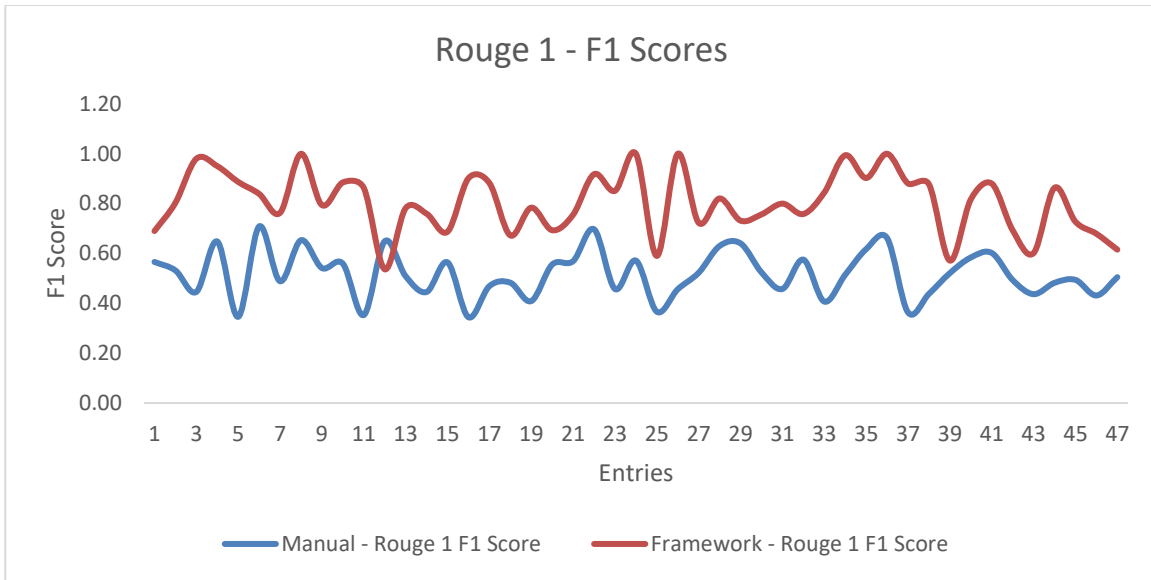


Figure 42 6.0 Comparison line graph for ROUGE-1 F1 score trend

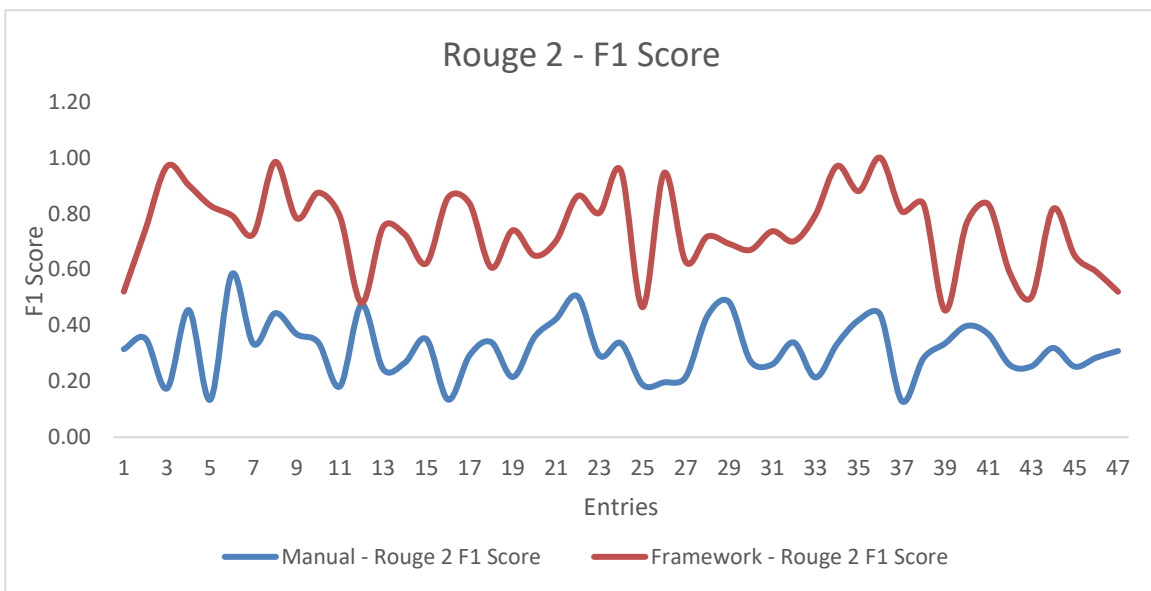


Figure 43 6.0 Comparison line graph for ROUGE-2 F1 score trend

The above graph shows the F1 score trend for each ROUGE-2, as we can visualize the F1 score trend for rouge-2 framework-generated summary is consistently higher than manual summary rouge-2 trend. Similarly, the next figure shows the same trend for F1 score, but for ROUGE-L.

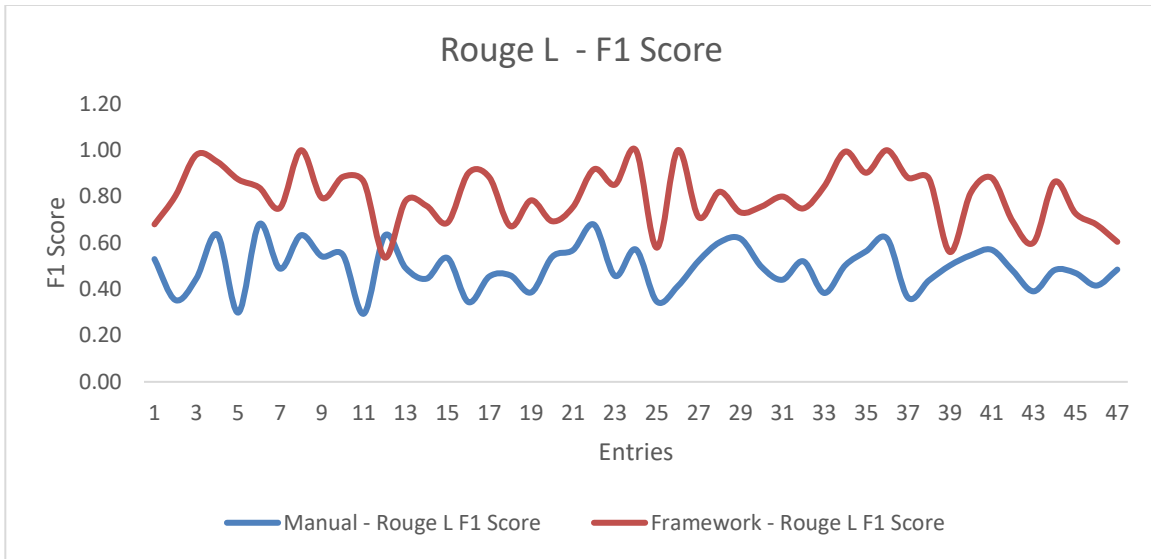


Figure 44 6.0 Comparison line graph for ROUGE-L F1 score trend

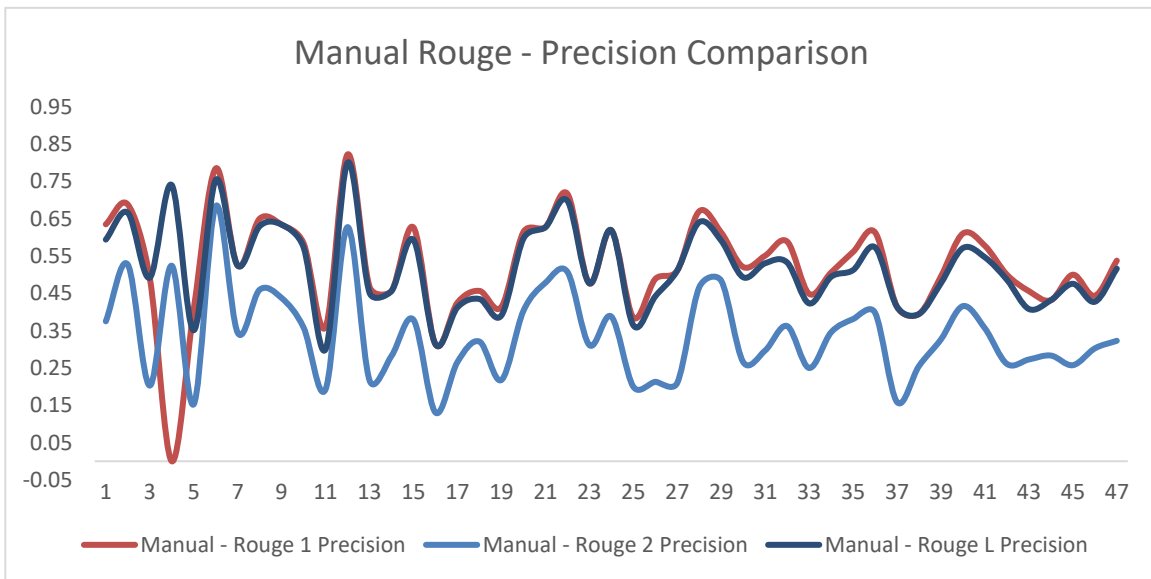


Figure 45 6.0 Precision trend comparison within manual summary Rouge scores

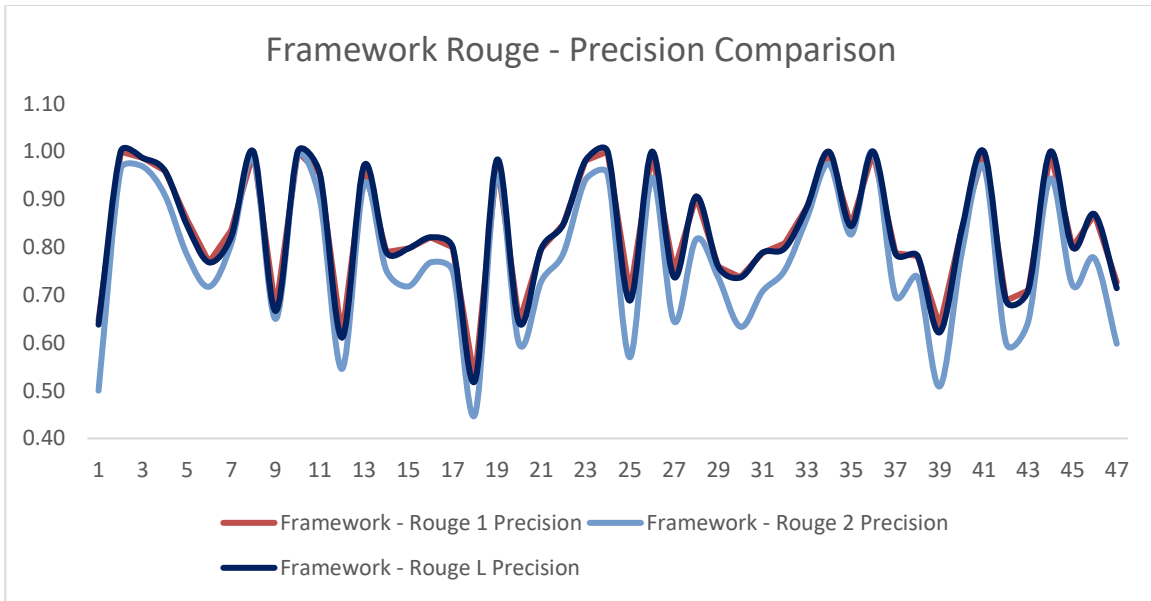


Figure 46 6.0 Precision trend comparison within framework summary Rouge scores

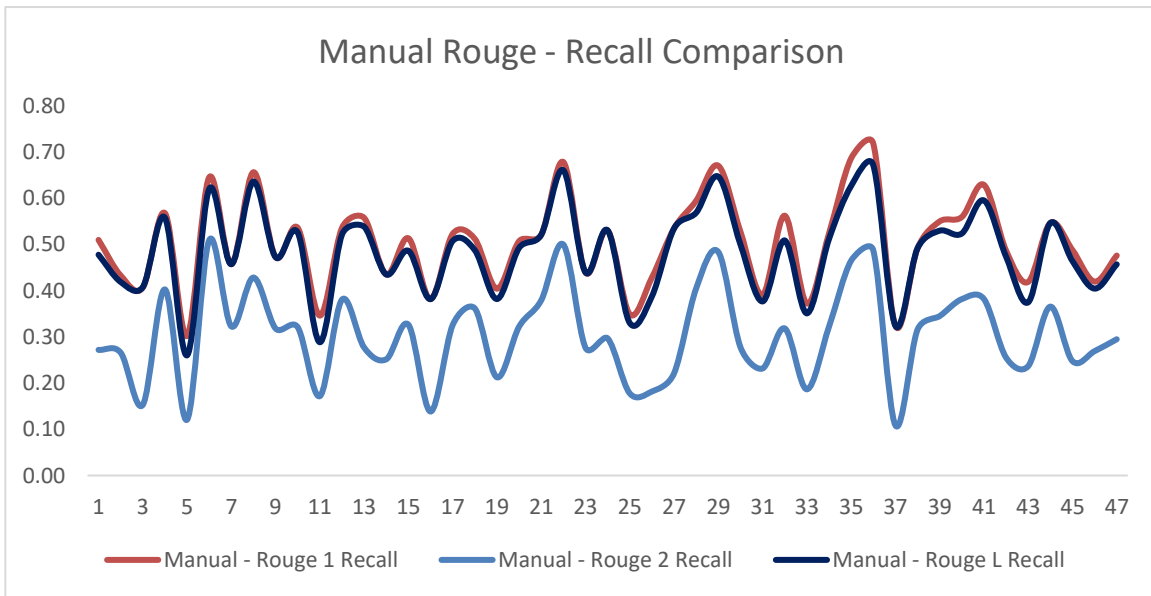


Figure 47 6.0 Recall trend comparison within manual summary Rouge scores

The above and below mentioned figures *figure 30-6.0* and *figure 31-6.0* displays the recall trend within the manual summary scores and framework summary's scores respectively.

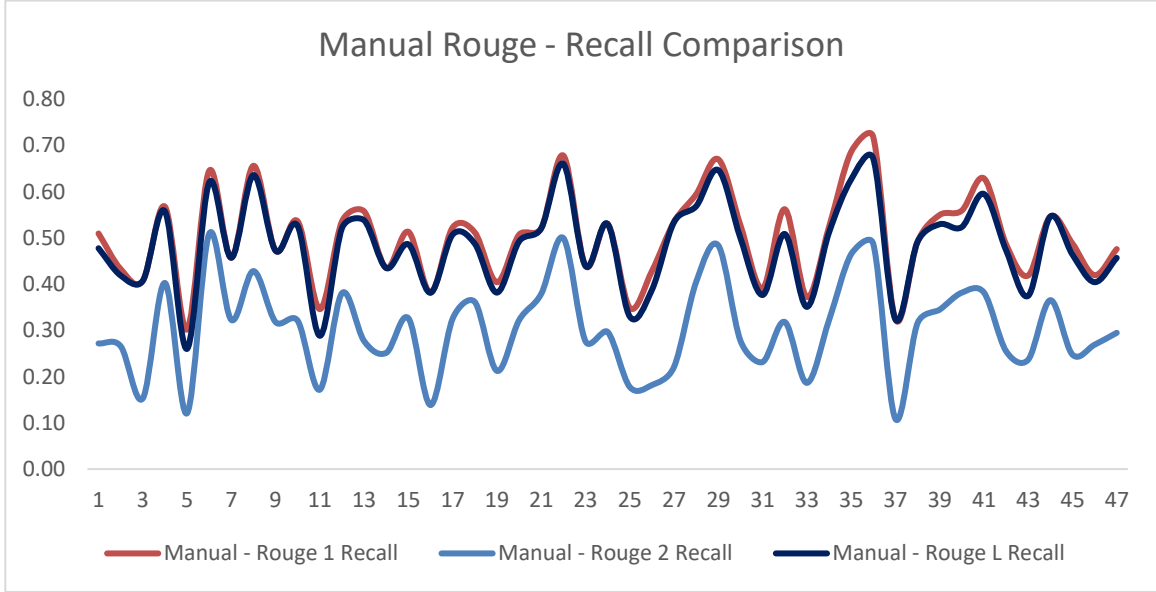


Figure 48 6.0 Recall trend comparison within manual summary Rouge scores

The above and below mentioned figures *figure 30-6.0* and *figure 31-6.0* displays the recall trend within the manual summary scores and framework summary’s scores respectively.

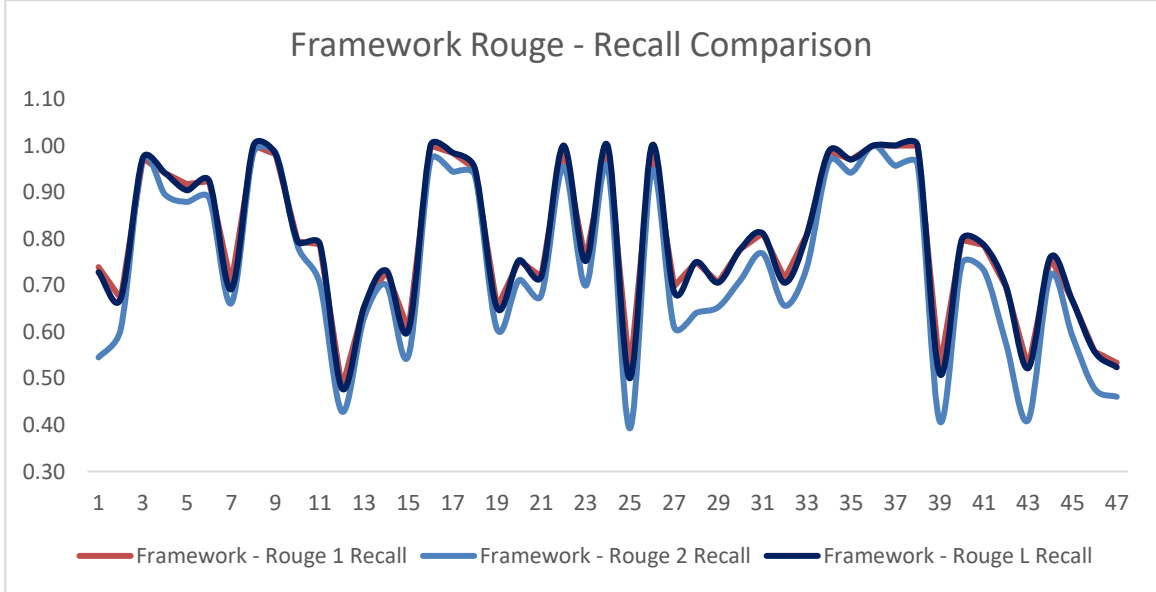


Figure 49 6.0 Recall trend comparison within framework summary Rouge scores

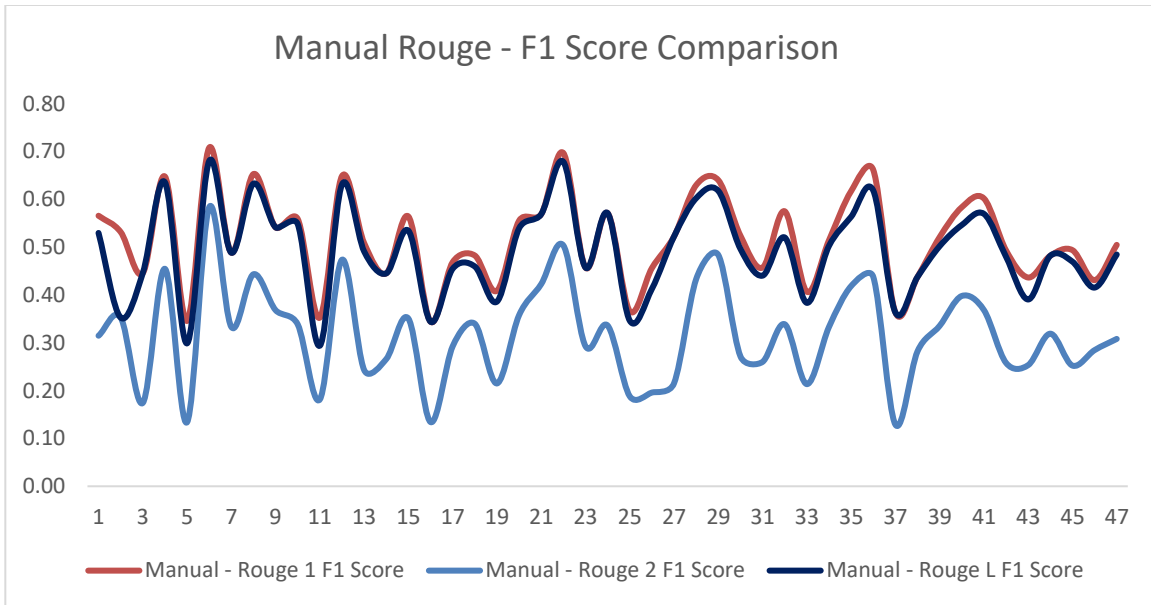


Figure 50 6.0 F1 trend comparison within manual summary scores

The above and below mentioned figures *figure 29-6.0* and *figure 30-6.0* displays the F1 trend within the manual summary scores and framework summary's scores respectively.

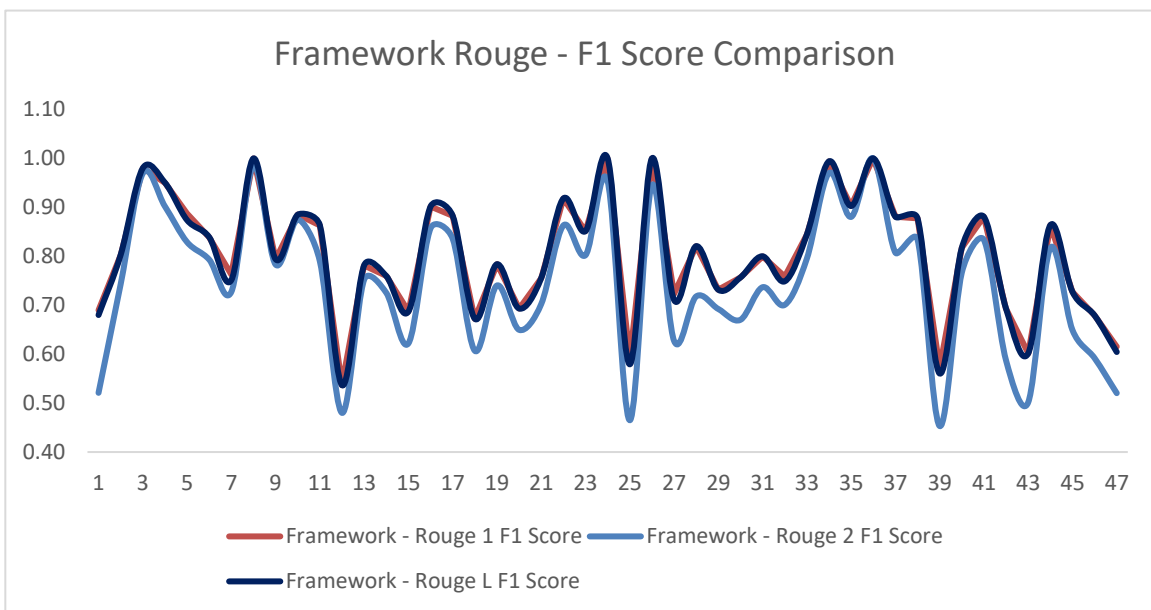


Figure 51 6.0 F1 trend comparison within framework summary scores

Upon compiling the above results and evaluating the final scores, it is evident that the summaries generated by our framework exhibits more relevance to the original reference summary. By putting the ROUGE scores in a graphical representation form exhibits a consistent trend, the framework-generated summary having the upward trend.

This trend underscores the effectiveness of our framework in producing summaries that closely aligns with the original reference summary. The higher trend of framework-generated summaries reflects its ability to capture and convey more accurate gist of the original content.

Chapter 7:

7. Conclusion and Future Work

In this section, we will conclude our work, and provide an insight into future work recommendations. In section 7.1 the conclusion of this research study is explained, and future work is explained in chapter 7.2.

7.1 Conclusions

Through a series of systemic steps, we embarked on a journey to develop a comprehensive approach for cross-lingual Urdu to English extractive text summarization. One of the key focuses of this research study was to leverage unsupervised learning techniques, instead of supervised learning. We opted for this approach in order to save time, resources and computational difficulties generally occur while using the supervised approach.

Our framework incorporated numerous components, including text translator, text summarization, summary evaluator through rouge score, and sentiment analysis for understanding the essence of the summarized text. Through this comprehensive framework, we aimed to address changes associated with cross lingual languages, especially the low resource language like Urdu. We successfully curated our dataset, using reference summaries as well as manually written summaries for the evaluation process.

Leveraging the translated limited resource language Urdu text to a high resource English language was the key decision made for using unsupervised approach. Due to the unpredictability and complexity of language patterns, we adopted the translation approach along with the unsupervised TextRank algorithm. By using this approach, we were able to avoid the need for large, annotated datasets.

For a thorough evaluation, we prepared dataset with rouge results for manually written summaries, in order to perform comparative analysis with the framework-generated summaries rouge scores the graphical analysis of rouge score comparison provided quantitative measures of summary quality. The trend displays a consistent high performance trend for framework-generated summaries as compared to the rouge scores of manual generated summaries. This trend shows the reliability and effectiveness of our framework in producing high quality summaries that closely aligned with the reference summaries and original text. In conclusion, our proposed approach is efficient and represents a significant contribution to the file of cross-lingual text summarization using unsupervised NLP techniques.

7.2Future Work

We can further advance the effectiveness and capabilities of our cross-lingual text summarization by applying alternative unsupervised algorithms, paving the way for comprehensive applications in diverse contexts and domains. Enhancing the performance by utilizing the potential of cross-lingual transfer learning techniques could be a way forward in future work.

Leveraging different language pairs and pre-trained language models can potentially improve the adaptability of our framework. Exploration of alternative NLP algorithms along with fine-tuning these models based on low resource language can provide novel performance enouncements.

Additionally, to capture the emotional nuances of original text in a better way, more advanced sentiment analysis techniques should be explored. Integration of multimodal data such as images, text and audio could enhance the comprehensiveness of the framework.

While our research has made significant progresses in the development of cross-lingual summarization framework using an unsupervised approach. By addressing these areas of future work, we can further advance the effectiveness of our framework.

REFERENCES

- [1] [2105.13648] Cross-Lingual Abstractive Summarization with Limited Parallel Resources (arxiv.org) [Online]. Available: <https://arxiv.org/abs/2105.13648>
- [2] [Jointly Learning to Align and Summarize for Neural Cross-Lingual Summarization] (<https://aclanthology.org/2020.acl-main.554>) (Cao et al., ACL 2020)
- [3] [Mixed-Lingual Pre-training for Cross-lingual Summarization] (<https://aclanthology.org/2020.aacl-main.53>) (Xu et al., AACL 2020)
- [4] Kaixiong Zhang, Yongbing Zhang, Zhengtao Yu, Yuxin Huang, Kaiwen Tan. A two-stage fine-tuning method for low-resource cross-lingual summarization[J]. *Mathematical Biosciences and Engineering*, 2024, 21(1): 1125-1143. doi: 10.3934/mbe.2024047
- [5] Rini Wijayanti, Masayu Leylia Khodra, Kridanto Surendro, Dwi H. Widyantoro, Learning bilingual word embedding for automatic text summarization in low resource language, *Journal of King Saud University - Computer and Information Sciences*
- [6] T. G. Varghese, C. V. Priya and S. M. Idicula, "A Novel Approach For English-Hindi Cross lingual Summarization," 2023 9th International Conference on Smart Computing and Communications (ICSCC), Kochi, Kerala, India, 2023, pp. 434-438, doi: 10.1109/ICSCC59169.2023.10335063. keywords: {Deep learning;Computational modeling;Transformers;Machine translation;Transformers;Text Summarization;Cross Lingual Summarization;Machine Translation},
- [7] Mohana, H. & Suriakala, Dr. (2021). An Enhanced Prospective Jaccard Similarity Measure (PJSM) to Calculate the User Similarity Score Set for E-Commerce Recommender System. 10.1007/978-981-15-5400-1_14.
- [8] Mahmood Yousefi-Azar, Len Hamey, Text summarization using unsupervised deep learning, *Expert Systems with Applications*
- [9] Priyadharshan, Thevatheepan & Sumathipala, Sagara. (2018). Text Summarization for Tamil Online Sports News Using NLP. 1-5. 10.1109/ICITR.2018.8736154.
- [10] S. A. A T, S. Shankaran, H. M. Thrupthi and M. H R, "Natural Language Processing based Cross Lingual Summarization," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2022, pp. 1825-1829, doi: 10.1109/ICOEI53556.2022.9776655. keywords: {Training;Predictive models;Market research;Mobile applications;Machine translation;Informatics;Testing;Natural Language processing;Abstractive Summarization;Extractive Summarization Translation;CLSTK;LSTM},
- [11] [Jointly Learning to Align and Summarize for Neural Cross-Lingual Summarization](<https://aclanthology.org/2020.acl-main.554>) (Cao et al., ACL 2020)
- [12] [2010.08892] Mixed-Lingual Pre-training for Cross-lingual Summarization (arxiv.org).

- [13] Pei-ying, Zhang & Cun-he, Li. (2009). Automatic text summarization based on sentences clustering and extraction. 167 - 170. 10.1109/ICCSIT.2009.5234971.
- [14] [Cross-lingual Fine-tuning for Abstractive Arabic Text Summarization](<https://aclanthology.org/2021.ranlp-1.74>) (Kahla et al., RANLP 2021)
- [15] Takeshita, S., Green, T., Friedrich, N. et al. Cross-lingual extreme summarization of scholarly documents. *Int J Digit Libr* 25, 249–271 (2024). <https://doi.org/10.1007/s00799-023-00373-2>
- [16] Z. Liu, J. Cao, J. Yu, G. Chen and Q. He, "ERNCLS: A Cross-Lingual Text Summarization Model Integrating Entity Relation Extraction Task," 2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), Chongqing, China, 2023, pp. 25-29, doi: 10.1109/ICIBA56860.2023.10165547.
- [17] S. A. A T, S. Shankaran, H. M. Thrupthi and M. H R, "Natural Language Processing based Cross Lingual Summarization," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2022, pp. 1825-1829, doi: 10.1109/ICOEI53556.2022.9776655.
- [18] S. Takeshita, T. Green, N. Friedrich, K. Eckert and S. P. Ponzetto, "X-SCITLDR: Cross-Lingual Extreme Summarization of Scholarly Documents," 2022 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Cologne, Germany, 2022, pp. 1-12.
- [19] M. Awais and R. Muhammad Adeel Nawab, "Abstractive Text Summarization for the Urdu Language: Data and Methods," in *IEEE Access*, vol. 12, pp. 61198-61210, 2024, doi: 10.1109/ACCESS.2024.3378300.
- [20] W. Bhatti and M. Aslam, "ISUTD: Intelligent System for Urdu Text De-Summarization," 2019 International Conference on Engineering and Emerging Technologies (ICEET), Lahore, Pakistan, 2019, pp. 1-5, doi: 10.1109/CEET1.2019.8711842.
- [21] N. Askarian, A. Fazly and A. Hamzeh, "A comparison of statistical measures for the automatic identification of Persian light verb constructions," *The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012)*, Shiraz, Iran, 2012, pp. 479-483, doi: 10.1109/AISP.2012.6313795.
- [22] K. D. Garg, V. Khullar and A. K. Agarwal, "Unsupervised Machine Learning Approach for Extractive Punjabi Text Summarization," 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2021, pp. 750-754, doi: 10.1109/SPIN52536.2021.9566038.
- [23] R. Habu, R. Ratnaparkhi, A. Askhedkar and S. Kulkarni, "A Hybrid Extractive-Abstractive Framework with Pre & Post-Processing Techniques To Enhance Text Summarization," 2023 13th International Conference on Advanced Computer Information Technologies (ACIT), Wrocław, Poland, 2023, pp. 529-533, doi: 10.1109/ACIT58437.2023.10275584.
- [24] S. S. Lwin and K. T. Nwet, "Extractive Summarization for Myanmar Language," 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Pattaya, Thailand, 2018, pp. 1-6, doi: 10.1109/iSAI-NLP.2018.8692976.
- [25] F. B and S. Abraham, "NLP Based Automated Business Report Summarization," 2022 International Conference on Innovative Trends in Information Technology (ICITIIT), Kottayam,

- India, 2022, pp. 1-4, doi: 10.1109/ICITIT54346.2022.9744151.
- [26] Y. Einieh and A. AlMansour, "Deep Learning in Arabic Text Summarization: Approaches, Datasets, and Evaluation Metrics," 2022 20th International Conference on Language Engineering (ESOLEC), Cairo, Egypt, 2022, pp. 45-49, doi: 10.1109/ESOLEC54569.2022.10009528.
- [27] P. B. Bafna and J. R. Saini, "Hindi Multi-document Word Cloud based Summarization through Unsupervised Learning," 2019 9th International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-19), Nagpur, India, 2019, pp. 1-7, doi: 10.1109/ICETET-SIP-1946815.2019.9092259.
- [28] K. Yamuna, V. Shriamrut, D. Singh, V. Gopaldasamy and V. Menon, "Bert-based Braille Summarization of Long Documents," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2021, pp. 1-6, doi: 10.1109/ICCCNT51525.2021.9579748.
- [29] A. W. Palliyali, M. A. Al-Khalifa, S. Farooq, J. Abinahed, A. Al-Ansari and A. Jaoua, "Comparative Study of Extractive Text Summarization Techniques," 2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA), Tangier, Morocco, 2021, pp. 1-6, doi: 10.1109/AICCSA53542.2021.9686867.
- [30] V. Dalal and L. Malik, "A Survey of Extractive and Abstractive Text Summarization Techniques," 2013 6th International Conference on Emerging Trends in Engineering and Technology, Nagpur, India, 2013, pp. 109-110, doi: 10.1109/ICETET.2013.31.
- [31] A. M. A. Danda, H. Bysani, R. P. Singh and S. Kanchan, "Automation of Text Summarization Using Hugging Face NLP," 2024 5th International Conference for Emerging Technology (INCET), Belgaum, India, 2024, pp. 1-7, doi: 10.1109/INCET61516.2024.10593316.
- [32] P. Raundale and H. Shekhar, "Analytical study of Text Summarization Techniques," 2021 Asian Conference on Innovation in Technology (ASIANCON), PUNE, India, 2021, pp. 1-4, doi: 10.1109/ASIANCON51346.2021.9544804.
- [33] B. Elayeb, A. Chouigui and M. Bounhas, "Analogical Text Mining: Application to Arabic Text Summarization and Classification," 2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA), Giza, Egypt, 2023, pp. 1-8, doi: 10.1109/AICCSA59173.2023.10479304.
- [34] S. K. Dam, M. Shirajum Munir, A. D. Raha, A. Adhikary, S. -B. Park and C. S. Hong, "RNN-based Text Summarization for Communication Cost Reduction: Toward a Semantic Communication," 2023 International Conference on Information Networking (ICOIN), Bangkok, Thailand, 2023, pp. 423-426, doi: 10.1109/ICOIN56518.2023.10048944.
- [35] M. Awais and R. Muhammad Adeel Nawab, "Abstractive Text Summarization for the Urdu Language: Data and Methods," in IEEE Access, vol. 12, pp. 61198-61210, 2024, doi: 10.1109/ACCESS.2024.3378300.
- [36] Rao, N.Srinivas. (2024). Text Summarization Based on Semantic Similarity. INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT. 08. 1-5. 10.55041/IJSREM32218.
- [37] Karthik, A.. (2024). Text and Video Summarization Using ML. International Journal for

Research in Applied Science and Engineering Technology. 12. 2577-2581. 10.22214/ijraset.2024.59243.

- [38] Urlana, Ashok. (2023). Enhancing Text Summarization for Indian Languages: Mono, Multi and Cross-lingual Approaches. 10.13140/RG.2.2.29544.24321.
- [39] Taspinar, Eymen & Yetis, Yusuf & Cihan, Onur. (2022). Abstractive Turkish Text Summarization and Cross-Lingual Summarization Using Transformer. 10.4018/978-1-6684-6001-6.ch011.
- [40] Pan, Hangyu & Xi, Yaoyi & Wang, Ling & Nan, Yu & Su, Zhizhong & Cao, Rong. (2023). Dataset construction method of cross-lingual summarization based on filtering and text augmentation. PeerJ Computer Science. 9. e1299. 10.7717/peerj-cs.1299.
- [41] Taunk, Dhaval & Sagare, Shivprasad & Patil, Anupam & Subramanian, Shivansh & Gupta, Manish & Varma, Vasudeva. (2023). XWikiGen: Cross-lingual Summarization for Encyclopedic Text Generation in Low Resource Languages.
- [42] Maurya, Kaushal & Desarkar, Maunendra & Kano, Yoshinobu & Deepshikha, Kumari. (2021). ZmBART: An Unsupervised Cross-lingual Transfer Framework for Language Generation.
- [43] Wang, Jiaan & Meng, Fandong & Zheng, Duo & Liang, Yunlong & Li, Zhixu & Qu, Jianfeng & Zhou, Jie. (2023). Towards Unifying Multi-Lingual and Cross-Lingual Summarization.
- [44] Shi, Xiaorui. (2023). MCLS: A Large-Scale Multimodal Cross-Lingual Summarization Dataset. 10.1007/978-981-99-6207-5_17.
- [45] Ma, Jiushun & Huang, Yuxin & Wang, Linqin & Huang, Xiang & Peng, Hao & Yu, Zhengtao & Yu, Philip. (2024). Augmenting Low-Resource Cross-Lingual Summarization with Progression-Grounded Training and Prompting. ACM Transactions on Asian and Low-Resource Language
- [46] Gupta, Pooja & Nigam, Swati & Singh, Rajiv. (2023). A Statistical Language Modeling Framework for Extractive Summarization of Text Documents. SN Computer Science. 4. 10.1007/s42979-023-02241-x.
- [47] Gambhir, Mahak & Gupta, Vishal. (2024). Improved hybrid text summarization system using deep contextualized embeddings and statistical features. Multimedia Tools and Applications. 1-30. 10.1007/s11042-024-19524-x.
- [48] Chhikara, Garima & Sharma, Anurag & Gurucharan, V. & Ghosh, Kripabandhu & Chakraborty, Abhijnan. (2024). LaMSUM: A Novel Framework for Extractive Summarization of User Generated Content using LLMs.
- [49] González Barba, José & Segarra, Encarna & García Granada, Fernando & Sanchis, Emilio & Hurtado Oliver, Lluís. (2023). Attentional Extractive Summarization. Applied Sciences. 13. 1458. 10.3390/app13031458.
- [50] Liu, Journey & Chen, Kuan-Yu & Hsieh, Yu-Lun & Chen, Berlin & Wang, Hsin-min & Yen, Hsu-Chun & Hsu, Wen-Lian. (2017). A Position-Aware Language Modeling Framework for Extractive Broadcast News Speech Summarization. ACM Transactions on Asian and Low-Resource Language Information Processing. 16. 1-13. 10.1145/3099472.

- [51] L. Guerrouj, D. Bourque and P. C. Rigby, "Leveraging Informal Documentation to Summarize Classes and Methods in Context," 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering, Florence, Italy, 2015, pp. 639-642, doi: 10.1109/ICSE.2015.212.
- [52] N. Vanetik, E. Podkaminer and M. Litvak, "Summarizing Financial Reports with Positional Language Model," 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 2023, pp. 2877-2883, doi: 10.1109/BigData59044.2023.10386704.
- [53] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 International Conference on Data Science and Communication (IconDSC), Bangalore, India, 2019, pp. 1-3, doi: 10.1109/IconDSC.2019.8817040.
- [54] A. Sharaff, A. S. Khaire and D. Sharma, "Analysing Fuzzy Based Approach for Extractive Text Summarization," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 2019, pp. 906-910, doi: 10.1109/ICCS45141.2019.9065722.
- [55] A. I. Seid, A. A. Abdisalan, M. M. Abdulahi, S. Parida and S. R. Dash, "Somali Extractive Text Summarization," 2022 OITS International Conference on Information Technology (OCIT), Bhubaneswar, India, 2022, pp. 1-6, doi: 10.1109/OCIT56763.2022.00063.
- [56] N. Pandey, S. Kumar, V. Ranjan, M. Ahamed and A. K. Sahoo, "Analyzing Extractive Text Summarization Techniques and Classification Algorithms: A Comparative Study," 2024 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC), Bhubaneswar, India, 2024, pp. 1-5, doi: 10.1109/ASSIC60049.2024.10508020.
- [57] N. Raychawdhary, N. Hughes, S. Bhattacharya, G. Dozier and C. D. Seals, "A Transformer-Based Language Model for Sentiment Classification and Cross-Linguistic Generalization: Empowering Low-Resource African Languages," 2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings), Mount Pleasant, MI, USA, 2023, pp. 1-5, doi: 10.1109/AIBThings58340.2023.10292494.
- [58] N. M. Magangane, S. G. Zwane and M. O. Adigun, "Datasets Collection Framework for Low-Resourced Languages in South Africa," 2024 Conference on Information Communications Technology and Society (ICTAS), Durban, South Africa, 2024, pp. 69-74, doi: 10.1109/ICTAS59620.2024.10507140.
- [59] X. Li, D. R. Mortensen, F. Metze and A. W. Black, "Multilingual Phonetic Dataset for Low Resource Speech Recognition," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 6958-6962, doi: 10.1109/ICASSP39728.2021.9413720.
- [60] J. -H. Wang, "Unsupervised multilingual concept discovery from daily online news extracts," 2010 IEEE International Conference on Intelligence and Security Informatics, Vancouver, BC, Canada, 2010, pp. 132-134, doi: 10.1109/ISI.2010.5484763.
- [61] Y. Chen and Q. Song, "News Text Summarization Method based on BART-TextRank Model," 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 2021, pp. 2005-2010, doi: 10.1109/IAEAC50856.2021.9390683.
- [62] D. Gunawan, S. H. Harahap and R. Fadillah Rahmat, "Multi-document Summarization by using TextRank and Maximal Marginal Relevance for Text in Bahasa Indonesia," 2019 International

- Conference on ICT for Smart Society (ICISS), Bandung, Indonesia, 2019, pp. 1-5, doi: 10.1109/ICISS48059.2019.8969785.
- [63] Y. Shi, B. Zhu and G. Li, "Research on automatic text summarization technology based on ALBERT-TextRank," 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Wuhan, China, 2022, pp. 841-844, doi: 10.1109/AEMCSE55572.2022.00169.
- [64] A. Rahman, F. M. Rafiq, R. Saha, R. Rafian and H. Arif, "Bengali Text Summarization using TextRank, Fuzzy C-Means and Aggregate Scoring methods," 2019 IEEE Region 10 Symposium (TENSYP), Kolkata, India, 2019, pp. 331-336, doi: 10.1109/TENSYP46218.2019.8971039.
- [65] N. Akhtar, M. M. S. Beg and H. Javed, "TextRank enhanced Topic Model for Query focussed Text Summarization," 2019 Twelfth International Conference on Contemporary Computing (IC3), Noida, India, 2019, pp. 1-6, doi: 10.1109/IC3.2019.8844939.
- [66] D. Bankira, S. Panda, S. Ranjan, H. S. Ali, S. Parida and N. Walubita, "Automatic Extractive text Summarization for Ho Language," 2023 OITS International Conference on Information Technology (OCIT), Raipur, India, 2023, pp. 915-919, doi: 10.1109/OCIT59427.2023.10430990.
- [67] S. R. Manalu, "Stop words in review summarization using TextRank," 2017 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Phuket, Thailand, 2017, pp. 846-849, doi: 10.1109/ECTICon.2017.8096371.
- [68] M. A. Hakim bin Sazali and N. B. Idris, "Neural Machine Translation for Malay Text Normalization using Synthetic Dataset," 2022 10th International Conference on Information and Communication Technology (ICoICT), Bandung, Indonesia, 2022, pp. 386-390, doi: 10.1109/ICoICT55009.2022.9914841.
- [69] W. S. A. Kurera, R. K. Rajapaksha, H. A. P. Rupasinghe, K. L. D. U. B. Liyanage, S. Kasthuriarachchi and S. Rajapakshe, "Building NLP Tools to Process Sinhala Text Data Written using English Letters," 2022 22nd International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 2022, pp. 063-068, doi: 10.1109/ICTer58063.2022.10024080.