

# Abstractive Text Summarization of Judicial Cases in Supreme Court of Pakistan



By

**Muneeb Ahmed Anwar**

**2017-NUST-MS-IT-18 205226**

Supervisor

**Dr. Faisal Shafait**

**Department of Computing**

A thesis submitted in partial fulfillment of the requirements for the degree  
of Masters of Science in Information Technology (MS IT)

In

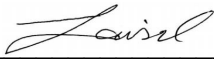
School of Electrical Engineering and Computer Science,  
National University of Sciences and Technology (NUST),  
Islamabad, Pakistan.

(August 2021)

## Approval

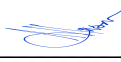
It is certified that the contents and form of the thesis entitled "Abstractive Text Summarization of Judicial Cases in Supreme Court of Pakistan" submitted by MUNEEB AHMED ANWAR have been found satisfactory for the requirement of the degree

Advisor : Prof. Dr. Faisal Shafait

Signature: 

Date: 11-Aug-2021

Committee Member 1:Mr. Adnan Ul-Hasan

Signature: 

Date: 12-Aug-2021

Committee Member 2:Dr. Muhammad Imran Malik

Signature: 

Date: 17-Aug-2021

Committee Member 3:Dr. Muhammad Shahzad

Signature: 

Date: 11-Aug-2021

# Dedication

I would like to dedicate this thesis to our Prophet Muhammad (S.A.W.W) for whom everything is created.

I would also like to dedicate this thesis to the pious family (Ahl-e-Bayt) of our Prophet Muhammad (S.A.W.W). Their piety and sacrifices are a great lesson for anyone and everyone. Next, I would like to dedicate this thesis to my parents who have always gone above and beyond to assure my utmost comfort. It is due of their prayers, love, and support that enabled me to reach where I am today.

I would also like to dedicate this thesis to all my teachers and mentors. Their dedication, affection and guidelines are the very reason I am at a good standing point in my life.

I would also like to dedicate this thesis to my wife for always being there to help and support me through all the thick and thin in various stages of life.

I would also like to dedicate this thesis to my kids, whose love and affection always kept me motivated to moving ahead in my life.


I would also like to dedicate this thesis to all of my family and friends who helped me, guided me, motivated me, and especially prayed for me.

And last but not the least, I would like to dedicate this thesis to anyone and everyone who believes in their goals and dreams.

## Certificate of Originality

I hereby declare that this submission titled "Abstractive Text Summarization of Judicial Cases in Supreme Court of Pakistan" is my own work. To the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics, which has been acknowledged. I also verified the originality of contents through plagiarism software.

Student Name: MUNEEB AHMED ANWAR

Student Signature: 

## THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis entitled "Abstractive Text Summarization of Judicial Cases in Supreme Court of Pakistan" written by MUNEEB AHMED ANWAR, (Registration No 00000205226), of SEecs has been vetted by the 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/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: 

Name of Advisor: Prof. Dr. Faisal Shafait

Date: 11-Aug-2021

Signature (HOD): \_\_\_\_\_

Date: \_\_\_\_\_

Signature (Dean/Principal): \_\_\_\_\_

Date: \_\_\_\_\_

# Acknowledgment

I would firstly like to pay my gratitude to Allah Almighty for showering upon me His countless blessings at every instance of life. Without His Will I couldn't have imagined completing this task.

I am most thankful to Dr. Faisal Shafait for encouraging and guiding me at every step to complete my thesis not only as a wonderful supervisor but as a brilliant mentor as well. His contribution in stimulating suggestions and encouragement, helped me to coordinate my thesis into its final form.

I am also very thankful to Dr. Adnan-ul-Hassan for his continuous guidance and support. And many thanks to Dr. Imran Malik and Dr. Muhammad Shehzad for joining us on this journey.

I am also very thankful to my father Mr. Muhammad Anwar, and my mothers, Ms. Shahida Khanum and Ms. Tahira Akhtar for helping me in making my dreams come true.

I am very thankful to my wife Ms. Madiha Rahim for always keeping me motivated and for always being on my side through all the thick and thin in life.

I am also very thankful to my kids Haider Ali, Muhammad Saad, Muhammad Arham, and Khadija Muneeb for bringing colors and motivation in my life.

There is a long list of brothers and sisters whom I would like to thank for

helping, guiding, motivating, and especially praying for me.

Special thanks to my brothers Abdullah Chohan, Naveed Iqbal, Abdul Samad, Ahmad Hassan, Ali Waqas, Haseeb Javed, Muhammad Usman and Maaz Hanafi.

Another special thanks goes to my sisters Ms. Muneeba Anwar, Ms. Adan Fatima, Ms. Nadia Kalsoom, Ms. Momna Saeed, Ms. Maah-e-Mubeen, Ms. Komal Fayyaz, Ms. Abida, Ms. Tahira, Ms. Nisar Fatima, Ms. Fariha Iqbal, Ms. Bushra Javed, Ms. Maria Alvi, and Ms. Rehana.

There is also a special group of people whom I would like to say thanks for their prayers and silent support that could not be seen but was felt. For that, I am very much thankful to Sir Hafeez Ahmed, Sir Maajid Maqbool, Sir Rasheed Ahmed, Sir Shahryar Khan, Mr. Ahmad Khalaf, Ms. Evarizza Quijano, and Sir Hassan Mehmood Zaidi.

And in the end, I am very much thankful to the School of Electrical Engineering and Computer Sciences department of Computing, and the faculty for invoking in me a strong educational foundation, which enabled me to complete this thesis.

# Table of Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Overview . . . . .	1
1.2	Natural Language Processing Using Deep Learning . . . . .	2
1.3	Problem Domain . . . . .	3
1.4	Problem Statement . . . . .	4
1.5	Solution Objective . . . . .	4
1.6	Solution Statement . . . . .	4
1.7	Thesis Organization . . . . .	5
1.7.1	Chapter 2: Literature Review . . . . .	5
1.7.2	Chapter 4: Methodology . . . . .	5
1.7.3	Chapter 5: Results and Discussion . . . . .	5
1.7.4	Chapter 6: Conclusion and Future work . . . . .	6
<b>2</b>	<b>Literature Review</b>	<b>7</b>
2.1	Area of research . . . . .	7
2.2	Extractive Text Summarization . . . . .	7
2.3	Abstractive Text Summarization . . . . .	10
2.4	Summarization of Legal Documents . . . . .	12
<b>3</b>	<b>Methodology and Dataset</b>	<b>15</b>
3.1	Methodology . . . . .	15
3.1.1	Base Model Selection . . . . .	15
3.1.2	Setup Base Model . . . . .	16
3.1.3	Base Model Evaluation . . . . .	16
3.1.4	Dataset Preprocessing . . . . .	16
3.1.5	Model Training . . . . .	17



## TABLE OF CONTENTS

3.1.6	Evaluation . . . . .	17
3.2	Dataset . . . . .	18
3.2.1	Data Preprocessing . . . . .	19
3.2.2	Dataset Split . . . . .	19
3.3	Base Model . . . . .	20
3.3.1	Abstractive text summarization Module (ATS) . . . . .	22
3.3.2	Neural text simplification Module (NTS) . . . . .	23
3.3.3	Dataset - Eureka Alert . . . . .	24
3.4	Evaluation . . . . .	24
3.4.1	System Evaluation . . . . .	25
3.4.2	Human Evaluation . . . . .	26
<b>4</b>	<b>Results and Discussion</b>	<b>27</b>
4.1	System Evaluation . . . . .	27
4.1.1	Text Summarization Evaluation . . . . .	27
4.1.2	Text Simplification Evaluation: . . . . .	31
4.1.3	Total (Summarization and Simplification) Evaluation: . . . . .	32
4.2	Summarization vs Simplification Evaluation: . . . . .	33
4.2.1	For $\beta = 0.0$ : . . . . .	34
4.2.2	For $\beta = 0.1$ : . . . . .	36
4.2.3	For $\beta = 0.2$ : . . . . .	38
4.2.4	For $\beta = 0.3$ : . . . . .	40
4.2.5	For $\beta = 0.5$ : . . . . .	42
4.2.6	For $\beta = 0.9$ : . . . . .	44
4.3	Aggregated Evaluation . . . . .	46
4.4	Human Evaluation: . . . . .	47
<b>5</b>	<b>Conclusion and Future Work</b>	<b>49</b>
5.1	Conclusion . . . . .	49
5.2	Future Work . . . . .	49

# List of Figures

3.1	Example Court Hearing . . . . .	18
3.2	Example Summary of a Case Hearing . . . . .	19
3.3	Final form of the Pre-Processed Dataset . . . . .	20
3.4	HTSS Layered Architecture . . . . .	22
3.5	HTSS layered architecture . . . . .	23
4.1	ROUGE-1 Evaluation . . . . .	28
4.2	ROUGE-2 Evaluation . . . . .	29
4.3	ROUGE-L Evaluation . . . . .	30
4.4	SARI Evaluation . . . . .	31
4.5	CSS1 Evaluation . . . . .	33
4.6	SARI, in relation with ROUGE I for $\beta = 0.0$ . . . . .	34
4.7	SARI, in relation with ROUGE II for $\beta = 0.0$ . . . . .	35
4.8	SARI, in relation with ROUGE L for $\beta = 0.0$ . . . . .	35
4.9	SARI, in relation with ROUGE I for $\beta = 0.1$ . . . . .	36
4.10	SARI, in relation with ROUGE II for $\beta = 0.1$ . . . . .	37
4.11	SARI, in relation with ROUGE L for $\beta = 0.1$ . . . . .	37
4.12	SARI, in relation with ROUGE I for $\beta = 0.2$ . . . . .	38
4.13	SARI, in relation with ROUGE II for $\beta = 0.2$ . . . . .	39
4.14	SARI, in relation with ROUGE L for $\beta = 0.2$ . . . . .	39
4.15	SARI, in relation with ROUGE I for $\beta = 0.3$ . . . . .	40
4.16	SARI, in relation with ROUGE II for $\beta = 0.3$ . . . . .	41
4.17	SARI, in relation with ROUGE L for $\beta = 0.3$ . . . . .	41
4.18	SARI, in relation with ROUGE I for $\beta = 0.5$ . . . . .	42
4.19	SARI, in relation with ROUGE II for $\beta = 0.5$ . . . . .	43

*LIST OF FIGURES*

4.20 SARI, in relation with ROUGE L for  $\beta = 0.5$  . . . . . 43  
4.21 SARI, in relation with ROUGE I for  $\beta = 0.9$  . . . . . 44  
4.22 SARI, in relation with ROUGE II for  $\beta = 0.9$  . . . . . 45  
4.23 SARI, in relation with ROUGE L for  $\beta = 0.9$  . . . . . 45  
4.24 Rouge-1, SARI and CSS1 scores in relation to  $\beta$  . . . . . 46  
4.25 Human-written Summaries vs System-generated Summaries . . . 47

# Abstract

Due to the large quantity of legal data availability on the internet, and other resources, it is vital for the research groups to carry broad research in the field of legal text processing, which can assist us make sense out of the huge quantity of obtainable data. This data expansion has forced the necessity to build systems that can help legal professionals as well as common citizens get important legal information with very little work. Legal document summarization is one of the most vital areas in legal domains. In this research, we apply and evaluate the performance of the hybrid text summarization and simplification (HTSS) algorithm on the dataset of court hearings of the Supreme Court of Pakistan. The results showed that the system-generated outlines are not very accurate despite having attained good scores from evaluation metrics like ROUGE, SARI and CSS.

**Keywords: Abstractive Text Summarization, Text Simplification, Legal, Judiciary, Court Hearings**

# Chapter 1

## Introduction

This chapter provides the opening and general information of the research to provide a clear understanding about this thesis. It covers the problem statement along with solution statement. It also describes the road map for our thesis and briefly highlights the further organization and structure of the thesis. Furthermore, it explains the motivation for carrying out the research work. Moreover, this chapter also gives idea about the vital contributions, scope of the work and key objectives of the thesis.

### 1.1 Overview

Extensive digitalization of official documents has been made in recent past. As per industrial research (Markets and Markets, 2018) market size of cloud storage will expected to be increased to \$89 Billion by year 2022. Usually, Adaption of technological frameworks for digital transformation in the field of legal industry appear very slothful.

This sector of industry mostly deals with critical and highly confidential official and personal documents. With the advancement in encryption standards and information assurance security, the digitalization of legal documents has enormously improved in legal sector which improves cost and time saving in legal sector.

However, judges require summarized and accurate information as simple digitalization of documents contains overload information for judgmental or-

ders. With the extensive digital documentation of a judgmental order, there is an emergent need to develop a system that can summarize the multiple documents with significantly lesser with accuracy and relevance.

Text simplification and text summarization are two separate tasks in natural language generation. Text simplification tends to decrease the complexity in a document, while text summarization tries to lessen the length of overall text at the same time keeping the original meaning of document intact.

Summarization is an assignment of compacting a portion of text into its briefer version, lessening the size of the initial text while instantaneously storing critical information and classifications of subject. As handwritten summaries are an expensive and tedious task, the growth of this work is gaining increasing popularity and is therefore a powerful reason for academic research.

There are important implications for text summaries in various NLP related activities such as answering questions, classifications, legal documents summarization, stories summarization, and making headlines. In addition, summaries can be integrated into these programs as an intermediate section that helps reduce text length.

## 1.2 Natural Language Processing Using Deep Learning

In natural language processing (NLP), summarization of text and simplification is a well-established mechanism. The main challenge in text summarization is to decrease the size of the document while preserve the appropriate information in source initial document.

In order to meet the desired requirement, are two types of broad approaches have been used which are extractive and abstractive methods. In Extractive method, exclusive summaries from whole passage are fetched from source text directly. However, in abstractive method, novel phrases and words which are not part of the source text have been generated in order to maintain the relevance of abstract with original document.

Implementation of extractive approach is simpler as accuracy and grammatically of paragraphs can be ensured while extracting large sets of text from source document. Whereas abstractive method uses sophisticated approach like paraphrasing, incorporation of real-world knowledge and generalization for high quality, accurate and relevant summarization.

The increase in computational power has enabled us to produce improved and more complex neural networks and deep learning models. This has opened doors to research and development in the Abstractive Text Summarization domain. Due to complex nature of abstractive summarization, majority of work had been done in the domain of extractive summarization methods. In recent past, sequence to sequence model was presented by Sutskever et al in 2014 which uses recurrent neural networks for read and generating the content has increase viability of abstractive summarization. Although these are promising systems; but they possess' inability to reproduce factual details correctly and unable to deal out of vocabulary words. Furthermore, most of the work observed in recent past focused on single sentence / headline generation work, while focus of our research is based on multi sentence summaries. To address challenging task of multi sentence summarization which requires higher level abstraction and avoid repetition, Machine learning and deep learning with natural language processing will be used.

### 1.3 Problem Domain

Judges in judicial sector face issues in processing extensive information in documents. Despite having digital documentation, the summaries in judgmental orders are processed manually which is very resource and time-consuming activity. With advancement in Natural Language processing, the same can be automated to speedup the judicial processing of court hearing.

## 1.4 Problem Statement

Manual summarization of judicial document requires specialized staff to summarize the case. The same not only requires time and resource but also requires legal analytics to avoid elimination of important and relevant information of the case. This will result in piling up and delaying of judicial process. Further, manual summarization can be biased and can eliminate the important findings intentionally or unintentionally which could change the decision of the case.

## 1.5 Solution Objective

We have provided a brief overview of objective of the solution to the earlier-mentioned problem in our thesis:

- Finding an efficient way of generating the judicial cases summary using Abstractive Text Summarization.
- Implementing the suitable approach to solve the above-mentioned problem.
- Applying the court-hearings dataset from Supreme Court of Pakistan.
- Evaluate the developed system using defined metrics such as, BLEU, ROUGE, and human evaluation.
- Verifying the accuracy of the system by comparing the system-generated summaries to the hand-written summaries.

The above statements define our solution objective briefly.

## 1.6 Solution Statement

Natural Language Processing (NLP) using deep learning has been used for automated summarization of digital documentation of judicial cases. Due to importance of accuracy and relevance of important facts, the automated summaries must be well appropriate and must contain important events /



facts to avoid ambiguity. The solution to address the peculiar requirement in judicial case summaries is to design a solution based on abstractive text summarization approach of NLP. We intend on generating simplified summaries for easier understanding. Therefore, we intend to use HTSS algorithm to generate summaries abstractively. HTSS was originally trained and used with the Eureka Science Alert dataset. HTSS will not only allow us to generate text summaries, but will also enable us to get simplified version of the summaries. Then we will evaluate the system using defined metrics (ROUGE for summarization, SARI for simplification, and CSS1 for for evaluating the combined task to summarization and simplification). We will also verify the results by comparing the system-generated summaries against the human-written summaries.

## **1.7 Thesis Organization**

Thesis organized in the following chapters:

### **1.7.1 Chapter 2: Literature Review**

This chapter explains the work done so far related to text summarization and legal text summarization techniques. It also provides Abstractive and Extractive existing approaches.

### **1.7.2 Chapter 4: Methodology**

This section provides a brief introduction to the simulation framework as well as a thorough description of the suggested algorithm and its features. Furthermore, the suggested algorithm's operation and implementation are described in depth.

### **1.7.3 Chapter 5: Results and Discussion**

This chapter will demonstrate the functioning and outcomes of our suggested system. The acquired simulation findings and their discussion round off the chapter.

#### **1.7.4 Chapter 6: Conclusion and Future work**

Brief summary of the thesis research work is presented in this section provided with tasks that can be carried out later for further research findings.

# Chapter 2

## Literature Review

*This chapter explains the related work done so far in the field of summarization of digital documents using NLP. The formulation of the thesis and the novelty of the thesis lie in identifying the research gap from the literature already published. The identification of the direction of research is also one of the sanctions of literature.*

### 2.1 Area of research

A thorough literature research was carried out in order to discover papers relating to existing summarizing techniques. According to research, there are several approaches accessible for text summarising. The supplied literature review is split into the sub-sections listed below. The first section contains papers in which academics created a summarising approach based on an extractive method, whereas the second subsection focuses only on abstractive techniques and the third paragraph summarises current strategies for legal document summarization.

### 2.2 Extractive Text Summarization

Current single document summarization methods established for news essays depends on a singular approach to summarize entire input documents. This is off great insignificance because high performance can not be achieved.

## CHAPTER 2. LITERATURE REVIEW

An article [1] proposed an Integer Linear Programming (ILP) approach by incorporating a modern regression-based methodology for single-document summarization.

The stated approach exclusively depends on a concept-based ILP technique to produce numerous candidate summaries for every input article investigating distinct concept weighting procedures and interpretation forms. Subsequently, a model augmented through numerous extracted features is incorporated sentence, n-gram, and summary level. The regression model is utilized for choosing amongst the candidates of the highly informative summary. This was established on an assessment of the conventional ROUGE-1 score. The examined characteristics were originated from the statistics of content significance, for example, position, coverage, and Frequency.

Neural Networks also offer effective assistance for extractive summarization of documents. An approach named TSRENN [2] fills the gap of redundancy and neglected relation between document and abstract caused by using Neural Networks. This is a two-tier approach. The first is RNN based extractive summarization following key sentence extraction. The sentence vector & Levenshtein distance is used hybrid sentence similarity measure in the extraction phase. The second phase consists of constructing GRU as basic blocks and demonstrating the document based on LDA to sustain summarization.

Usually, the automated function for summarizing text comprises of spontaneously compiling a document to offer a shortened form of it. Producing a summary involves not only the assortment of key topics but also the detection of vital relationships between the topics. Correlated tasks place text units, especially sentences, to choose those that can form a summary. Though, the consequential summary may not consist of all topics in the source text because crucial info may have been squandered.

Supplementary, computer-generated text documents have not been analyzed extensively in this field. Hence, a study [3] proposed a modern technique of automatic text summarization (ATS) function that utilizes semantic information to enhance keyword detection. The suggested approach not only enriches coverage by combining sentences to detect key topics in the core text but also precision in detecting key words in paragraphs.

Greatest text specifics on the Internet makes automated text summarizations extremely essential these days. Instantly, the objective of multitasking documentation is to obtain summaries from document compendium, and at the same time to cover key content and to reduce needless information. Though, various methodologies to various objectives have earned significance because their results have enhanced those that have the same persistence.

On the other side, in multi-objective use, the modified schemes have generated encouraging findings in other applications. For this reason, an algorithm based on the Multi-Objective Artificial Bee Colony (IMOABC) [4] algorithm has been developed and used in the process of shortening the text of multiple texts. Experimentations were accomplished on data from the Document Understanding Conferences (DUC) data set, and the outcomes were compared with the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric. The findings improved to those in the scientific literature between 7.37% and 40.76% and 2.59% and 11.24% of ROUGE-2 and ROUGE-L, respectively.

The evident growth of Web documents has necessitated the automation of documents summarization. In this perspective, a summary of the document being released, that is, the task of extracting the most appropriate information eliminates inefficiency and offers the remaining details jointly and confidentially structure, it is a challenging task. Another article [5] proposed a clever approach named ExDoS, earning the advantages of both supervised and unsupervised learning at the same time.

ExDoS is the first way to integrate both supervised and unsupervised algorithms into one Outline and interpretive approach to summarize documents. ExDoS iteratively downs the error rate in each collection with the help of dynamic local element measurement. In addition, this approach clarifies the provision of features to discriminate against each class, which is a difficult issue in the summarization mission. Therefore, in adding to summarizing the text, ExDoS is also able to scale the file and the importance of each element in the summarization procedure.

Thanks to the vast amount of data available today, text summarization has become increasingly important to find the right amount of information in large print. We look at extensive articles on blogs, news websites, customer

review websites, and so on. A review article [6] offers a variety of ways to summarize key texts. Several papers have been discussed in various ways that have been used so far to summarize the text.

In particular, the techniques portrayed in the article generate Abstractive (ABS) or Extractive (EXT) text summaries. Summarization strategies based on problems are also discussed. The article examines structured based and semantic methods of summarizing texts. Several datasets have been applied to test the simulations made by these genres, such as DUC2000, single and multi-documentation, CNN corpus, etc.

## 2.3 Abstractive Text Summarization

This paper [7] discusses the recent findings in automatic text summarization focusing the summarization systems based on neural networks. It is very important to improve the design of existing automatic summarization systems so that the requirements of increasing data can be handled. This paper presents a overview of many neural network based abstractive summarization models.

The proposed framework consists of five key parts named as encoder-decoder architecture, optimization algorithms, training techniques, dataset, mechanisms, and evaluation metric. This study provides a wide understanding of the parts of recent neural networks based abstractive text summarization models. Based on the analysis, models using a transformer-based encoder-decoder architecture are considered more advanced. For abstractive summarization, this research recommends using the pre-trained language models in balance with neural network architecture. This paper provides the design patterns of the latest abstractive summarization systems.

Also, this study discusses the different types of languages and mechanisms used for the models of abstractive systems. This paper highlights certain gaps like using the pretrained models of “MASS” (Masked sequence to sequence pre-training for language generation) and “BART” for summarization systems. The design elements discussed in this study are considered helpful for the implementation of novel abstractive summarization system.

A well-defined goal of Artificial Intelligence is to develop an abstractive

text summarization (ATS) system which can generate relevant and accurate summaries of documents. The significant impacts of using the deep learnings schemes have benefited the ATS systems.

This [8] paper proposes a unique Hierarchical Human-like deep neural network for ATS (HH-ATS), stimulated by the process of how human users understand an article and write the relevant summary. This proposed framework is comprised of three main parts named as knowledge-aware hierarchical attention module, a multitask learning module, and a dual discriminator generative adversarial network (DD-GAN). This framework represents the three stages of human reading understanding (casual reading, attentive reading, and postediting). Experimental results show that HH-ATS significantly surpasses the compared methods as it attains higher ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores than advanced standard techniques. Also, study indicates that Hierarchical Human-like deep neural network for ATS (HH-ATS) can generate summaries with improved fluency.

This paper [9] discusses an automatic text summarization (ATS) model, that expands conventional sequence- to-sequence (Seq2Seq) neural text summarization technique by implementing a syntax-augmented encoder and a headline-aware decoder. The encoder utilizes both syntactic structure and word knowledge of a sentence in the sentence construction. Syntactic units are focused by proposed hierarchical attention mechanism.

A headline attention mechanism and a Dual-memory-cell LSTM network is used to enhance the decoder to get increased quality of generated summaries. To compare the proposed model with other advanced models, experiments were performed on CNN/DM datasets. These experiments show that the proposed model provides the better results as compared to existing models and achieve a summary generation equivalent to the extractive methods. The results produced by the proposed method shows the generation of more concise and less redundant summaries of source documents. So, this analysis presents the results as intended for proposed framework. In future directions, this paper also recommends including the designing of a flexible data structure to assist syntactic parsing trees for batch training.

## 2.4 Summarization of Legal Documents

A sentence's verbal status in a court document reveals phrase's aim regardless of whether it is a claim or hold up verification, it is advantageous to court document processing systems e.g., document retrieval systems in courts. Following approaches can assist the court in quickly summarising legal matters with accurate and trustworthy findings.

Deep learning has been demonstrated to be useful for natural language processing applications like conversation analysis. They [10] suggested tackling the issue of recognising the rhetorical state of each phrase in a court document with deep learning models. Also proposed that an artificial neural module embedded with rhetorical information might be useful for retrieval of information. Take the simplest way to modelling sentence dependency: use inter-sentence and inter-status relationship modelling. Use a recurrent neural network to represent inter-sentence relationships. Use conditional random field (CRF) technique to reproducing a reliance in inter-status connection.

Authors [11] introduce "JudgeDoll," a technology that automatically extracts the essential information from a lengthy judgement and produces a legal summary for Thai Supreme Court criminal and civil cases. The finite-state template matching algorithm is used to extract information. In terms of lexicons and stylistics, the linguistic patterns employed in legal documents, such as verdicts and testimony, are restricted. In an algorithm, the complete verdict's texts are first anchored as textual bodies. Then, using a preset set of regular expressions, capture the information of each textual body.

The authors of [12] work present an ontology-driven knowledge block summarization technique to compute document similarity. The experiment was carried out for Chinese judgement document categorization. First, from the views of top-level and domain related ontology and extra semantic knowledge for Chinese judgement papers is incorporated. where it is further illustrated how to combine various types of ontology in an extendable manner. Second, using ontology-based information extraction the essential semantic knowledge contained in papers may be summarised into knowledge blocks. Third, assess the similarity between various knowledge blocks using Word Mover's Distance (WMD). Finally, the KNN-based tests were carried out.



Frequently used extractive automated summarizing techniques Recall-Oriented Understudy for Gisting Evaluation metrics (ROUGE) and Bilingual Evaluation Understudy metrics (BLEU) comprehensively compare is in [13] work. The analysis is done on a publicly accessible data set. According to the findings of the experiments, graph oriented summarization methods better in assessment measures and additional crucial contextual information might help automated summarization algorithms to perform better.

This comparative analytical study may be used to provide a baseline for the benchmark legal data set, which will be useful for future research in this area. The [14] suggested approach employs a Seq2Seq Architecture using RNN to accomplish document summarising tasks. The summary based on abstractive technique, allowing the model to generate internal meaning on its own. After cleaning the data set, the algorithm removes stop words and punctuation before training with a predefined set of rules on trained data using TensorFlow and the seq-2-seq architecture. After training, test the model by producing summaries and calculate the consistency and ROUGE scores of the summaries.

The authors [15] offer an annotation method based on observations of the judicial system in Japan. They carried out a pilot research to assess the inter-annotator relationship. The experiment was carried out with the participation of two annotators. Extremely complicated sentence structure and technical vocabulary characterise the legal language employed in legal documents. FRAMING by Hachey and Grover was broken into two parts: FRAMING-main denotes the higher layers of the argumentation framework, whereas FRAMING-sub denotes the lower levels. The division enables for different levels of relevance of the supporting information covered by FRAMING to be distinguished. This [16] article provides a methodology for identifying the paragraphs of a case document that contribute to its summary and using those paragraphs to find other papers that are comparable. A Support Vector Classifier was trained using a data set including clearly labelled summary paragraphs of Indian Supreme Court papers. When opposed to evaluating the document as a whole, using simply the extracted summary to obtain comparable documents performs better in terms of time and space complexity.

## *CHAPTER 2. LITERATURE REVIEW*

To capture the primary clause inside a single document, the authors [17] employ the latent semantic analysis (LSA) approach. Employ an untrained method for a single civil document and a trained summary technique for multiple criminal documents, depending on the kind of input case. The data was gathered from legitimate government websites, including the Supreme Court.

# Chapter 3

## Methodology and Dataset

*This chapter provides a brief overview of Dataset, proposed architectures, followed by an evaluation techniques.*

### 3.1 Methodology

#### 3.1.1 Base Model Selection

Since the intent of this research was to find a way to generate simplified summaries abtractively on the case hearings dataset of Supreme Court of Pakistan, therefor we worked on identifying a base model that successfully performs Abtractive Text Summarization. Our primary preference was to identify a model that already performs summarization in legal domain. The idea was to identify such a model, and then fine-tune it (further improve its output), to not only fulfil our requirement, but also to improve the standard abtractive-legal text summarization domain. But unfortunately we could not find any abtractive text summarization model that was specifically summarizing court hearings.

Therefore, we went on to identifying and evaluating various other abtractive text summarization models. Our search lead us to HTSS: Hybrid Text Summarization and Simplification model. proposed by F. Zaman et al. (2020) [18]. We chose HTSS to be our base model because it not only outperformed other state-of-the-art abtractive text summarization models,

but also took on the task of simplifying the generated text summaries. More details about HTSS could be found in the Base Model section.

### 3.1.2 Setup Base Model

We first setup the base model and perform a manual evaluation of initial results. The model needed to be upgraded to use the latest versions of Python and other software packages. The artificial intelligence landscape progresses fast, so many improvements were to be made to source code of the original model since the release of the paper by Zaman et al. (2020) [18]. The original experiment was carried out on a system with Linux OS and Titan 1080 GPUs. The source code was written in Python.

However, we setup our experiment on Google Colab with Tesla 4 GPU and 12 GB RAM.

### 3.1.3 Base Model Evaluation

After successful setup of our Base Model, we first carried out the original experiment to assess the execution of the base model. This was needed to make sure that the experiment was producing the same output as mentioned in the paper [18]. Our base model evaluation yielded the results similar to those mentioned in the research paper. Therefor we were ready to proceed further with HTSS [18] as our base model.

### 3.1.4 Dataset Preprocessing

After successful selection, execution and evaluation of our base model, we preprocessed the dataset to make it compatible with the base model. Dataset preprocessing involved following steps:

**Dataset Cleansing:** This involved in modifying original hearing and summaries by removing any special characters, and extra spaces and white-lines.

**CSV Compilation:** We compiled the dataset in CSV format from the given

text files. The CSV contained document indices, ground truth (hand-written summaries), and relative paths to the original hearing text files.

**Vocabulary Generation:** refers to generating the unique vocabulary words from the original hearings and human written summaries.

**Dataset Split** This involved splitting the dataset into train dataset and test dataset.

More details about the dataset preprocessing could be found in the Dataset Section.

### 3.1.5 Model Training

Once the dataset was preprocessed and compiled, we fed the training dataset to the base model and generated trained models. The need for having different models arose because we had a hyperparameter  $\beta$  that needed to be fine-tuned to get the most optimal output.

Therefore we trained our model several times with different values of  $\beta$  to get the most optimal output.

### 3.1.6 Evaluation

We evaluated the performance of our model with test dataset after training the model with different values of hyperparameter  $\beta$ . We assessed the relevance and readability of our generated summaries using various evaluation metrics and also compared our generated summaries with reference summaries.

Below is the list of evaluation metrics and methods used for evaluating the model:

- ROUGE-1
- ROUGE-2
- ROUGE-L
- SARI

## CHAPTER 3. METHODOLOGY AND DATASET

- CSS1
- Human Evaluation

More details about the model evaluation metrics and methods are mentioned in the Evaluation section. The findings of our evaluation are mentioned in the next chapter.

### 3.2 Dataset

As the title of this research suggests, we have used Supreme Court of Pakistan’s case hearings as our dataset and human-generated case summaries as our ground truth. The dataset (hearings and summaries) contained 606 case hearings and summaries for the cases held in Supreme Court of Pakistan. The dataset we received was composed of 1212 raw text files. A single case was represented with 2 text files, one file containing the actual case hearing, and another file contained the human-generated summary of the corresponding case hearing.

```
P L D 2016 Supreme Court 64
Present: Anwar Zaheer Jamali, C.J., Amir Hani Muslim and Umar Ata Bandial, JJ
Sheikh MUHAMMAD ILYAS AHMED and others—Appellants
Versus
PAKISTAN through Secretary Ministry of Defence, Islamabad and others—Respondents
Civil Appeals Nos. 1125 and 1126 of 2014, decided on 29th October, 2015.
(On appeal from judgment of Lahore High Court, Rawalpindi Bench dated 18-4-2014, passed in RFAs Nos.144 and 145 of
2003, respectively)
Date of hearing: 29-10-2015.
JUDGMENT
ANWAR ZAHEER JAMALI, C J.—For the reasons set out in the applications for condonation of delay, the delay in filing
of these appeals is condoned and the appeals are taken up for hearing on merit.
2. At the outset, learned ASC for the appellants has made a statement at the bar that in view of announcement of
judgment by this Court today in connected Civil Appeals Nos.1120 to 1124 of 2014, the appellants are not pressing
these appeals for seeking further enhancement in the amount of compensation, but only to the extent of non awarding
of interest on the amount of compensation, as mandated under section 34 of the Land Acquisition Act, 1894 (in short
"the Act of 1894"), which has been withheld for no valid reasons.
3. A bare reading of above referred provision of the Act of 1894 reveals that awarding of such interest is statutory
in nature, which cannot be withheld. Thus, the appellants are fully entitled for grant of compound interest at the
rate of eight percent per annum from the date of taking possession of acquired land till the date of payment of its
compensation, but for no valid reasons, such relief has escaped the sight of the two Courts below.
4. This being the position, these appeals are partly allowed to the extent that the appellants will also be entitled
for compound interest at the rate of eight percent per annum from the date when possession of the acquired land was
taken over from them till the time, compensation in terms of the impugned judgment dated 18.4.2014, is paid to them.
HMA/M-47/S Order accordingly.
```

Figure 3.1: Example Court Hearing

### 3.2.1 Data Preprocessing

As the data was in a raw format, therefore it required some preprocessing before it could be used with the baseline model. The model accepts the dataset in a csv format as an input; containing the record id, Summary and file path to the original document (a case hearing in our case).

```
P L D 2016 Supreme Court 64

Present: Anwar Zaheer Jamali, C.J., Amir Hani Muslim and Umar Ata Bandial, JJ

Sheikh MUHAMMAD ILYAS AHMED and others—Appellants

Versus

PAKISTAN through Secretary Ministry of Defence, Islamabad and others—Respondents

Civil Appeals Nos. 1125 and 1126 of 2014, decided on 29th October, 2015.

(On appeal from judgment of Lahore High Court, Rawalpindi Bench dated 10-4-2014, passed in RFAs Nos.144 and 145 of 2003, respectively)

Land Acquisition Act (I of 1894)—

—S. 34—Payment of interest on compensation—Award of such interest was statutory in nature, and could not be withheld.

Altaf Elahi Sheikh, Sr. ASC for Appellants.

Sohail Mehmood, DAG. and Sqd. Ldr. Farhat Rafiq for Federation.

Date of hearing: 29-10-2015.
```

Figure 3.2: Example Summary of a Case Hearing

In the figure 3.3, id is a unique number assigned to each record, Text-Simplified is the human-generated summary for a particular case hearing, and File-Path is the relative path to the text file containing the actual case hearing text.

### 3.2.2 Dataset Split

The dataset was then split into a train set and test set by 80-20 ratio. Therefore, out of 606 samples, our training set contained 484 samples (80%) and

## CHAPTER 3. METHODOLOGY AND DATASET

our testing set contained 122 samples (20%). There was no need to have a validation set firstly because of the limited number of samples, and secondly because there is primarily one tunable hyper-parameter in HTSS. Once we pre-processed the data and split it into train and test subsets, we passed the dataset into the HTSS algorithm and generated the trained models. The HTSS algorithm primarily contains one hyper-parameter  $\beta$ . We trained various models with different values of  $\beta$  (0.0, 0.1, 0.2, 0.3, 0.5, 0.9). We then applied the test sets on the trained model to evaluate each model’s performance.

Id	Text_Simplified	File_Path
0	<p>2017 P T D 1481                      [Supreme Court of Pakistan]                      Present: Man Saqib Nisar, C.J., Umar Ata Bandial and Maqbool Baqar, JJ                      FEDERATION OF PAKISTAN through Secretary Revenue Division, Islamabad and others                      Versus                      Messrs SAHB JEE and others                      Civil Appeal No.1074 of 2008, decided on 19th January, 2017.                      (Against the judgment dated 20-3-2009 of the Lahore High Court, Lahore passed in W.P. No.11983 of 2005).                      (a) Establishment of Office of Federal Tax Ombudsman Ordinance (200XV of 2000)                      -Ss. 11 &amp; 32 Reference to the President against recommendation of Ombudsman Scope When the Revenue Division or any person was aggrieved of a recommendation made                      (b) Establishment of Office of Federal Tax Ombudsman Ordinance (200XV of 2000)                      -Ss. 11 &amp; 32 Reference to the President against recommendation of Ombudsman Scope Section 32 of the Ordinance, providing for representation before the President, did not                      (c) Establishment of Office of Federal Tax Ombudsman Ordinance (200XV of 2000)                      -S.14(B) Federal Tax Ombudsman Review, power of Scope Section 14(B) of the Establishment of Office of Federal Tax Ombudsman Ordinance, 2000 empowered the Ombudsman                      (d) Establishment of Office of Federal Tax Ombudsman Ordinance (200XV of 2000)                      -Ss. 14(B) &amp; 32 Reference to the President against order of Ombudsman passed in review jurisdiction Scope Whilst exercising power of review, if the Ombudsman sets aside                      Khalid Abbas Khan, Advocate Supreme Court for Appellants.                      Nemo for Respondent No.1.                      Ex parte for Respondents Nos.2 and 3.</p>	dataset/SC/descriptions/001.txt
1	<p>2015 S C M R 365                      [Supreme Court of Pakistan]                      Present: Nasser-ul-Mulk, C.J., Gulzar Ahmed and Muzir Alam, JJ                      CIVIL APPEALS NOS.1122 AND 1123 OF 2011                      (Against judgment dated 5-5-2011 of Federal Service Tribunal, Islamabad, passed in Appeal No.33(JCS of 2008)                      MUHAMMAD ZAFAR ALI and others Appellants                      Versus                      ASIM GULZAR and others Respondents                      CIVIL APPEAL NO.1343 OF 2014                      (Against order dated 3-10-2014 of High Court of Sindh at Karachi, passed in C.P. No. D-1085 of 2013)                      Syed MUHAMMAD ABBAS RIZVI and others Appellants                      Versus                      FEDERATION OF PAKISTAN and others Respondents                      CRIMINAL APPEAL NO. 436 OF 2011                      (Against order dated 9-8-2011 of High Court of Sindh, Circuit Court, Hyderabad, passed in C.P. No. D-198 of 2009)                      ASIM GULZAR and others Appellants                      Versus                      ATTAULLAH KHAN CHANDIO and others Respondents                      CIVIL APPEAL NO. 431 OF 2013                      (Against order dated 18-1-2013 of High Court of Sindh at Karachi, passed in C.P. No. D-3657 of 2009)                      ASIM GULZAR and others Appellants                      Versus                      ATTAULLAH KHAN CHANDIO and others Respondents                      Civil Appeals Nos. 1199, 1191 of 2011, Civil Appeal No. 1363 of 2014, Original Appeal No.236 of 2011 and Civil Appeal No.291 of 2013, decided on 16th December 2016.</p>	dataset/SC/descriptions/002.txt

Figure 3.3: Final form of the Pre-Processed Dataset

### 3.3 Base Model

We chose HTSS: Hybrid Text Summarization and Simplification model (proposed by F. Zaman et al. (2020) [18]) as our base model. It implements the Pointer-Generator Network proposed by [19]. (2017) with improved loss



### CHAPTER 3. METHODOLOGY AND DATASET

function. The reason for choosing this model is that it successfully generates simplified summaries that outperform standard abstractive text summarization models.

Simplification and summarization of the content are associated, but the sub-tasks are different in generating natural language. The summarization of text attempts to decrease the size of the document with sustaining the actual description of text, and simplification it makes it easier to try to reduce the complexity of the text. Number of international platforms of Science and humanities use these text summarization and simplification techniques in manual or automated ways. One of them is the well known Science platform 'Eureka Alert'. The source documents usually contain domain specific and hard to understand words for a common man, so it must be worked on to simplify that text to be understood.

A novel hybrid architecture is proposed by [18] to combine the task of summarization and simplification of abstractive and extractive summarization known as HTSS. For the task of combined simplification and summarization, [18] has extended the pointer generator network model. A corpus of five thousand plus Eureka Alert articles was attained and then a loss function for the Hybrid task was introduced. A binary score for hard and easy words are used in the look-up table [18].

In literature, three methods for the purpose of text simplification are syntactic simplification, lexical simplification and neural text simplification.

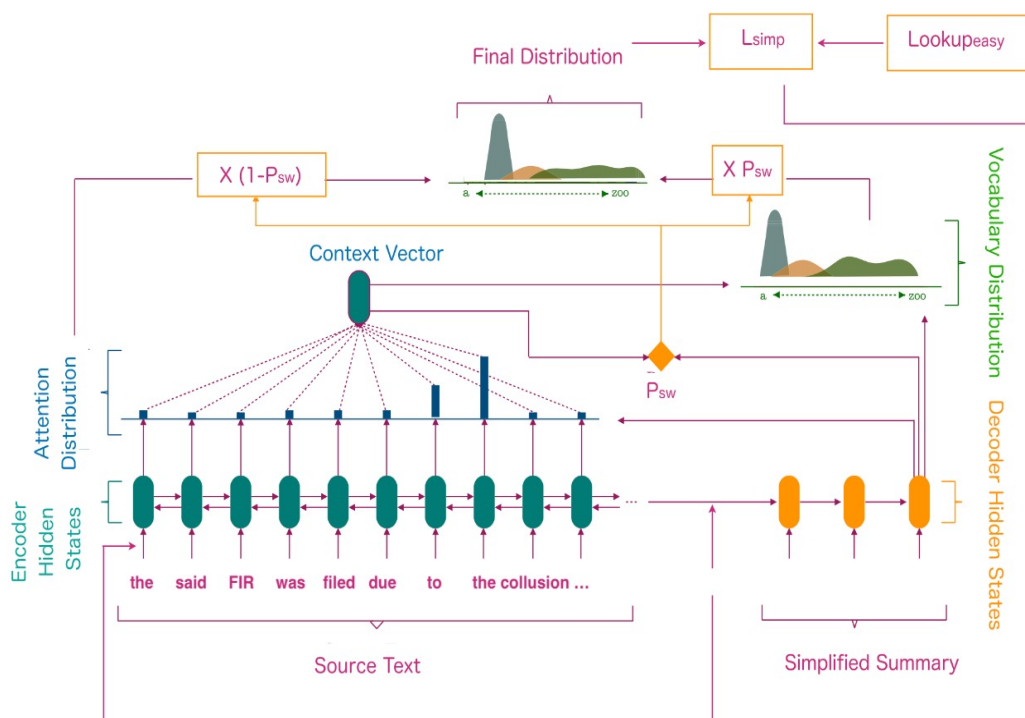


Figure 3.4: HTSS Layered Architecture

### 3.3.1 Abstractive text summarization Module (ATS)

Abstractive text summarization makes use of deep neural based sequence to sequence networks to generate summarizations on basis of provided text as input. ATS methods tend to generate long sections of text but the drawback of ATS is they generate non-factual sentences in output. ATS [20] is a data driven, end to end sequence to sequence model comprising of two levels made up of LSTM encoders with 1000 hidden units and word inserting with 500 dimensions.

Pointer generation model is presented in this section. The pointer-generator network is a hybrid between the sequence-to-sequence attention model (Nalapaty et al. (2016) [21]), and a pointer network (Vinyals et al., (2015) [22]), as it permits both copying words via pointing, and generating words from a fixed vocabulary. The model is best suited for simplification of text and summarization. The pointer generation model consist of a soft switch, bi

directional single layer encoder and a uni directional single layer LSTM. The decoder produces newly generated word with probability  $\epsilon \in [0, 1]$  from vocabulary distribution. A number of hidden states is produced by encoder by extracting the words from source input document. At the time of training, the decoder produce summary word using hidden state sequence which was produce by encoder and earlier word from reference summary at time step. At training time, the earlier word comes from the reference summary, but in case of testing, the word comes from the decoder.

### 3.3.2 Neural text simplification Module (NTS)

NTS [23], is a encoder-decoder sequence to sequence model having beam search. The encoder is made up of two LSTM layers with 500 hidden units, whereas the decoder contains two layers of LSTM with global attention.

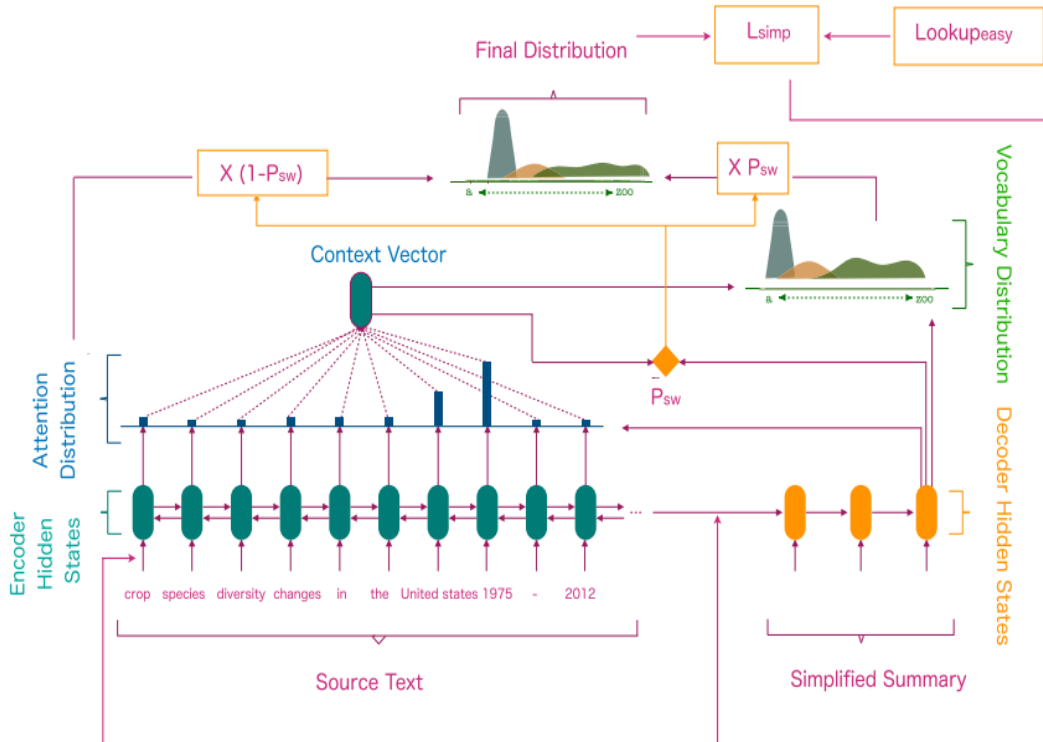


Figure 3.5: HTSS layered architecture

The pointer-generator network is a hybrid between our baseline and a

pointer network (Vinyals et al., 2015), as it permits both copying words by pointing, and generating words from a fixed vocabulary. Proposed HTSS abstractive simplified architecture Pointer generator model is expanded with an enhanced loss function, for the mutual text simplification and summarization, the function is further customized.

The easiness score is measured for summaries and the simplification loss is computed. It is done to impose the model to learn generating the shortened summary texts.

### 3.3.3 Dataset - Eureka Alert

The data for the task of simplification and summarization was necessarily collected as parallel quantity of summaries from the Eureka Alert website and subsequent scientific journal articles. Eureka alert, an online source where bloggers, researchers and other domain specialists take scientific contents and generate a summarized and readable text manually.

The summary is accessible to the public through the website of Eureka Alert. The task was to first extract 227,590 simplified and easy to read versions of the list of scholarly articles available at Eureka Alert. The summarised document is linked to its original article where its published with the help of DOI (Document Object Identifier). The DOIs provided access to get PDFs, and parsing the PDFs resulted in errors, so a filter was applied to fetch the articles which had an xml version.

So, only 5204 articles with summary wise pairs for following journals i.e. PLOS-ONE, Nature Communication, and Scientific Reports. The data was characterized based on the six attributes: Eureka Title Simplified, Eureka Text Simplified, Full Paper XML, Paper Title, Paper Journal and Paper DOI.

## 3.4 Evaluation

Evaluation methods of the output generated by the system can be broadly divided into 2 categories.

1. System Evaluation

2. Human Evaluation

**3.4.1 System Evaluation**

System evaluation refers to the set of software tools or metrics that evaluate the performance of a summary. Below are the evaluation metrics used to assess the output summaries generated by our system:

**ROUGE-1:** ROUGE-1 is a text summarization evaluation metric that refers to the overlap of each word between the system-generated summary and human-generated summary.

**ROUGE-2:** ROUGE-2 is a text summarization evaluation metric that refers to the overlap of pair of consecutive words between the system-generated summary and human-generated summary.

**ROUGE-L:** ROUGE-L is a text summarization evaluation metric that refers to the overlap of longest common sub sequence of consecutive words between the system-generated summary and human-generated summary.

**SARI:** SARI is a text simplification evaluation metric that compares system-generated simplified summary against the ground truth summary (human-generated summary) and at the same time against the system-generated non-simplified summary at a sentence level.

**CSS1:** CSS1 is an evaluation metric that evaluates the performance of a combined task of text summarization and simplification. CSS1 is actually a harmonic mean between ROUGE-1 and SARI scores. Below is the formula of CSS1.

$$CSS1 = \frac{2R_1XSARI}{R_1 + SARI} \quad (3.1)$$

### **3.4.2 Human Evaluation**

Human evaluation is the simplest and most commonly used evaluation method. It is also a necessity to get an output evaluated by the human beings, especially the domain experts and other stakeholders, so that the correctness of the system could be verified.

# Chapter 4

## Results and Discussion

The trained models were tested using the test dataset containing 122 samples. We evaluated the output of our generated summaries upon various metrics. The summaries were generated on trained models with different values of our hyper-parameter  $\beta$ .

The evaluation can be broadly distributed in two categories.

### 4.1 System Evaluation

System evaluation can be further classified in following categories:

- Text Summarization Evaluation
- Text Simplification Evaluation
- Total Evaluation

#### 4.1.1 Text Summarization Evaluation

The Text Summarization task was evaluated using ROUGE-1, ROUGE-2, and ROUGE-L metrics. Below are the outputs of each of the evaluation metric with respect to different values of  $\beta$ .

**ROUGE-1 Evaluation:**

From ROUGE-1 evaluation, it can be seen that the model performs slightly well for  $\beta = 0.1$ .

Below is the table describing the ROUGE-1 evaluation of generated summaries upon models with different values of  $\beta$ .

<b>Rogue-1</b>			
$\beta$	<b>Precision</b>	<b>Recall</b>	<b>F1-Score</b>
0	84.1	39.42	53.33
0.1	84.42	44.3	57.73
0.2	81.81	28.89	42.2
0.3	81.53	39.05	52.47
0.5	84.38	42.44	56.18
0.9	81.33	41.52	54.56

Table 4.1: ROUGE-1 Evaluation

The graph below describes the output in terms of ROUGE-1 evaluation for different values of beta. It can be seen from the graph that model performs well for  $\beta = 0.1$ , and performs worst for  $\beta = 0.2$ .

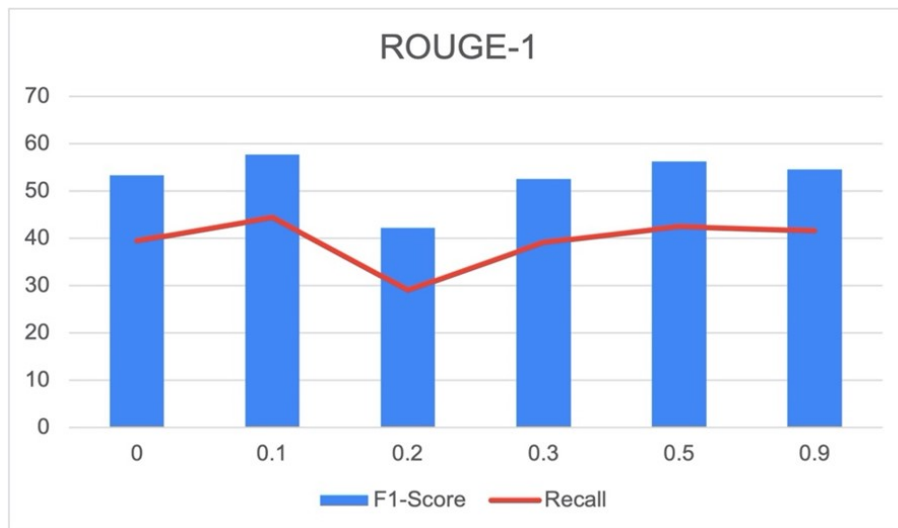


Figure 4.1: ROUGE-1 Evaluation



**ROUGE-2 Evaluation:**

Just like ROUGE-1, it can be seen that the model performs slightly well for  $\beta = 0.1$  for ROUGE-2 evaluation.

Below is the table describing the ROUGE-2 evaluation of generated summaries.

<b>Rogue-2</b>			
$\beta$	<b>Precision</b>	<b>Recall</b>	<b>F1-Score</b>
0	39.27	23.24	28.97
0.1	46.58	30.16	36.35
0.2	25.26	12.22	16.36
0.3	40.31	24.11	29.96
0.5	44.28	27.37	33.64
0.9	41.19	25.72	31.43

Table 4.2: ROUGE-2 Evaluation

The graph below describes the output in terms of ROUGE-2 evaluation for different values of beta. It can be seen from the graph that model performs well for  $\beta = 0.1$ , and performs worst for  $\beta = 0.2$ .

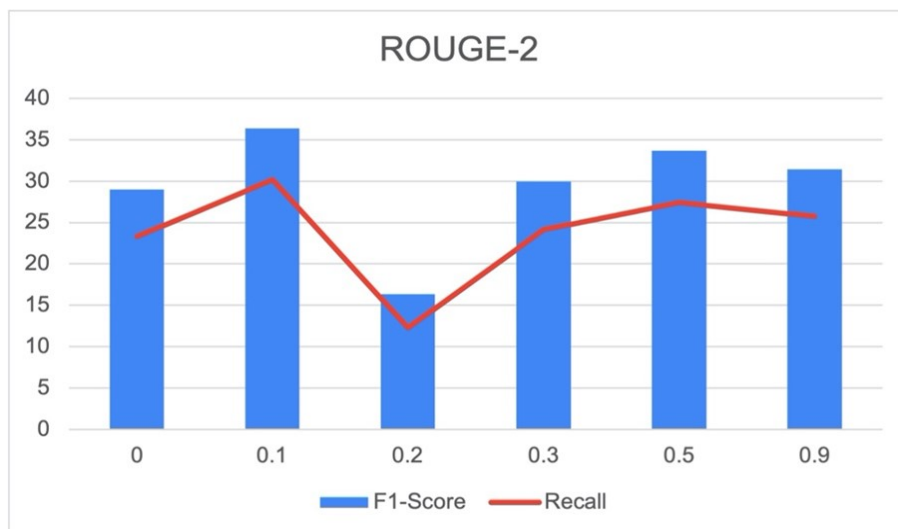


Figure 4.2: ROUGE-2 Evaluation

**ROUGE-L Evaluation:**

From ROUGE-L evaluation, it can be seen that the model performs slightly well for  $\beta = 0.1$ .

Just like it did for ROUGE-1 and ROUGE-2 evaluations. Below is the table describing the ROUGE-L evaluation of generated summaries.

<b>Rogue-L</b>			
$\beta$	<b>Precision</b>	<b>Recall</b>	<b>F1-Score</b>
0	83.79	39.28	53.14
0.1	84.16	44.29	57.55
0.2	80.21	28.55	41.71
0.3	81.22	38.88	52.25
0.5	84.29	42.17	55.84
0.9	81.19	41.45	54.47

Table 4.3: ROUGE-L Evaluation

The graph below describes the output in terms of ROUGE-L evaluation for different values of beta. It can be seen from the graph that model performs well for  $\beta = 0.1$ , and performs worst for  $\beta = 0.2$ .

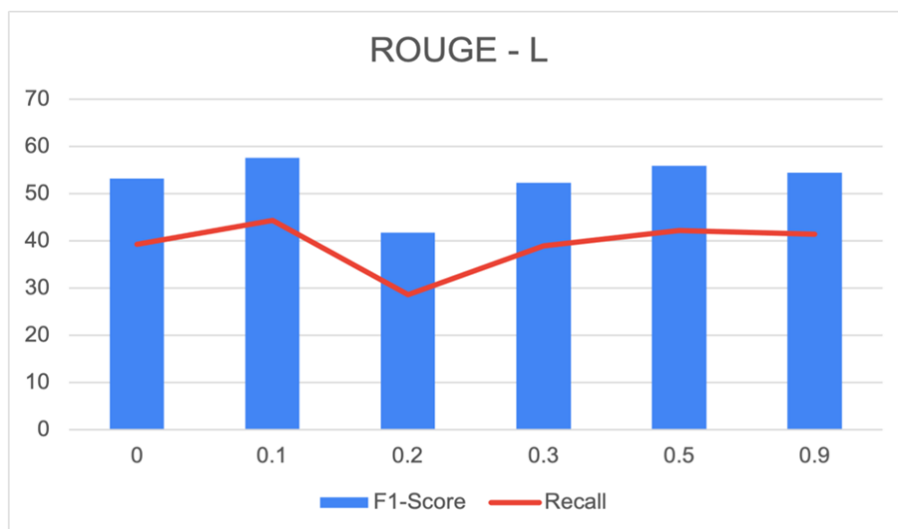


Figure 4.3: ROUGE-L Evaluation

### 4.1.2 Text Simplification Evaluation:

#### SARI Evaluation:

We used SARI for evaluating the simplified text in output summary. This evaluation also suggests that model performs slightly well for  $\beta = 0.1$ .

Just like it did for ROUGE-1, ROUGE-2 and ROUGE-L evaluations. Below is the table describing the SARI evaluation of generated summaries.

$\beta$	SARI
0	31.24
0.1	38.41
0.2	23.38
0.3	32.77
0.5	34.85
0.9	33.36

Table 4.4: SARI Evaluation

The graph below describes the output in terms of SARI evaluation for different values of beta. It can be seen from the graph that model performs well for  $\beta = 0.1$ , and performs worst for  $\beta = 0.2$ .

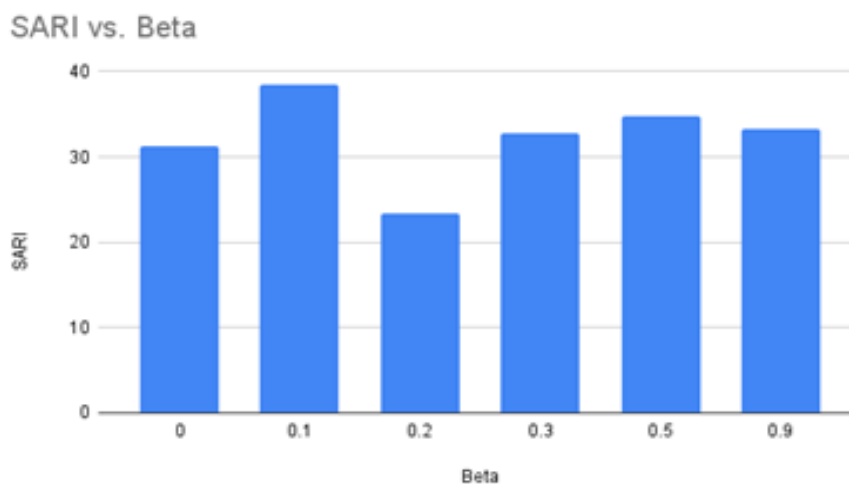


Figure 4.4: SARI Evaluation

### 4.1.3 Total (Summarization and Simplification) Evaluation:

Since no previous work was carried out in the past that takes care of text summarization and simplification simultaneously, therefore [18] proposed a new evaluation metric CSS1 for evaluating the combined task of text summarization and simplification. CSS1 is actually a harmonic mean between ROUGE-1 and SARI results.

#### CSS1 Evaluation:

We used CSS1 for evaluating the combined task of summarization and simplification in the output summary. This evaluation also suggests that model performs slightly well for  $\beta = 0.1$ .

Just like it did for ROUGE-1, ROUGE-2, ROUGE-L, and SARI evaluations. Below is the table describing the CSS1 evaluation of generated summaries.

$\beta$	CSS1
0	39.4
0.1	46.13
0.2	30.09
0.3	40.34
0.5	43.02
0.9	41.4

Table 4.5: CSS1 Evaluation

The graph below describes the output in terms of CSS1 evaluation for different values of beta. It can be seen from the graph that model performs well for  $\beta = 0.1$ , and performs worst for  $\beta = 0.2$ .

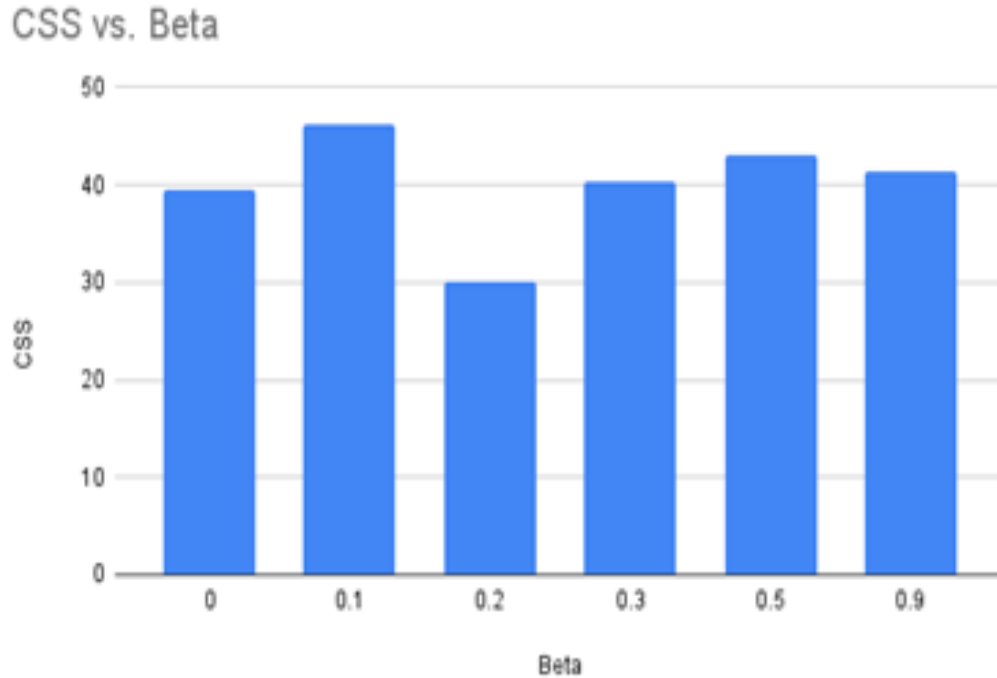


Figure 4.5: CSS1 Evaluation

## 4.2 Summarization vs Simplification Evaluation:

Next, we evaluate the output of SARI in relation to ROUGE-1, ROUGE-2 and ROUGE-L scores for different values of  $\beta$ .

$\beta$  is the hyper-parameter that decides how much attention we should give to the simplification task with respect to the summarization task.

We have shown the distribution of our data instances according to their SARI and ROUGE scores as scattered charts.

### 4.2.1 For $\beta = 0.0$ :

#### SARI vs ROUGE-1

The figure below is a scattered chart between SARI and ROUGE-1 for  $\beta = 0.0$ . It can be seen from the graph below that majority of the scores are centered around the mean values of SARI (31.24) and ROUGE-1 (53.33), despite having several outliers in the cluster. The mean values of SARI and ROUGE-1 scores are mentioned in Table 4.6.

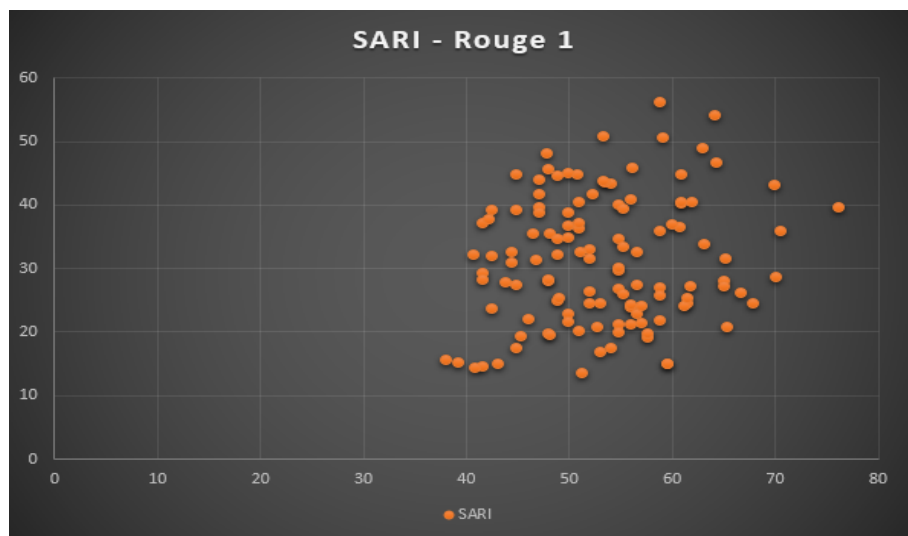


Figure 4.6: SARI, in relation with ROUGE I for  $\beta = 0.0$

#### SARI vs ROUGE-2

The figure below is a scattered chart between SARI and ROUGE-2 for  $\beta = 0.0$ . It can be seen from the graph below that majority of the scores are centered around the mean values of SARI (31.24) and ROUGE-2 (28.97), despite having several outliers in the cluster. The mean values of SARI and ROUGE-2 scores are mentioned in Table 4.6.

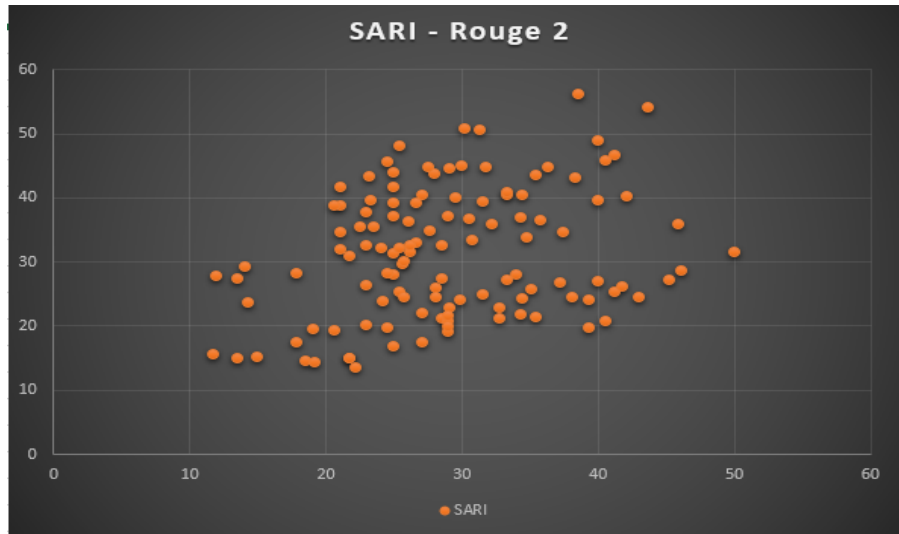


Figure 4.7: SARI, in relation with ROUGE II for  $\beta = 0.0$

### SARI vs ROUGE-L

The figure below is a scattered chart between SARI and ROUGE-L for  $\beta = 0.0$ . It can be seen from the graph below that majority of the scores are centered around the mean values of SARI (31.24) and ROUGE-L (53.17), despite having several outliers in the cluster. The mean values of SARI and ROUGE-L scores are mentioned in Table 4.6.

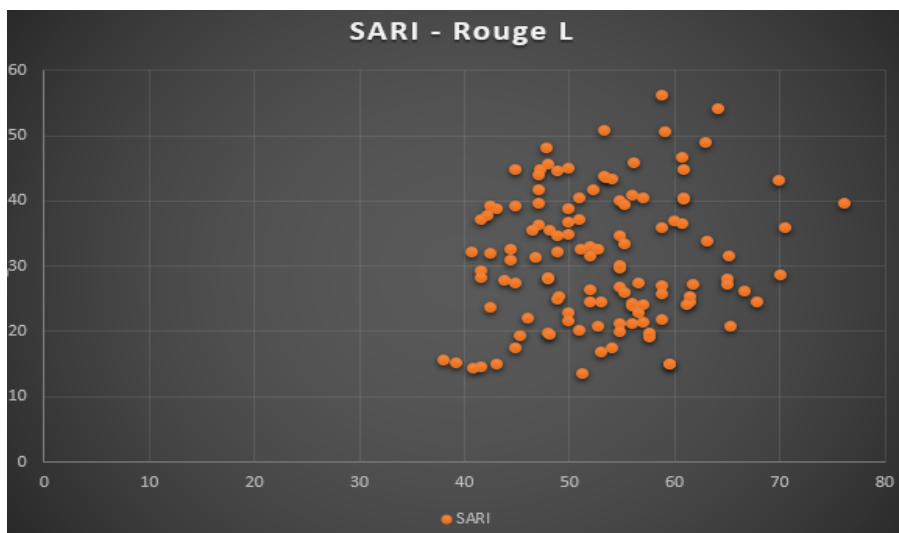


Figure 4.8: SARI, in relation with ROUGE L for  $\beta = 0.0$

### 4.2.2 For $\beta = 0.1$ :

#### SARI vs ROUGE-1

The figure below is a scattered chart between SARI and ROUGE-1 for  $\beta = 0.1$ . It can be seen from the graph below that majority of the scores are centered around the mean values of SARI (38.41) and ROUGE-1 (57.73), despite having several outliers in the cluster. The mean values of SARI and ROUGE-1 scores are mentioned in Table 4.6.

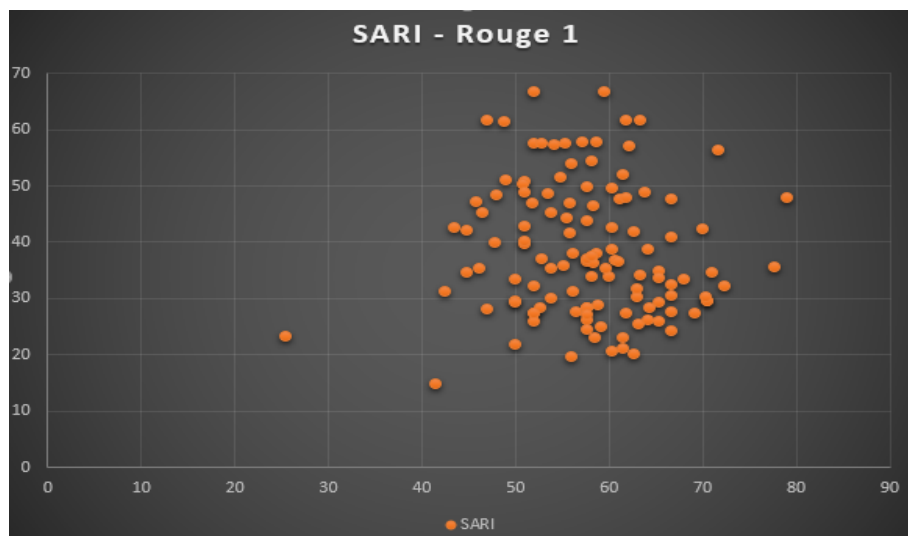


Figure 4.9: SARI, in relation with ROUGE I for  $\beta = 0.1$

#### SARI vs ROUGE-2

The figure below is a scattered chart between SARI and ROUGE-2 for  $\beta = 0.1$ . It can be seen from the graph below that majority of the scores are centered around the mean values of SARI (38.41) and ROUGE-2 (36.35), despite having several outliers in the cluster. The mean values of SARI and ROUGE-2 scores are mentioned in Table 4.6.



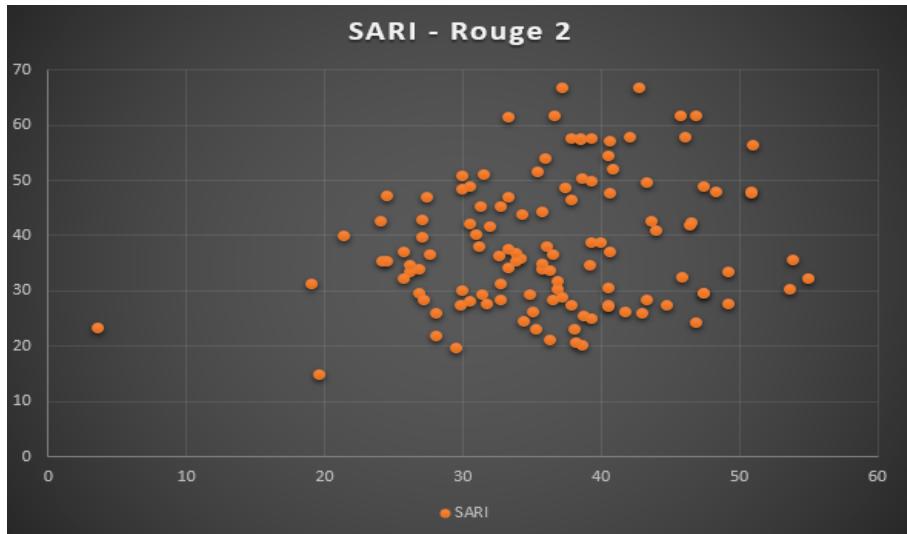


Figure 4.10: SARI, in relation with ROUGE II for  $\beta = 0.1$

### SARI vs ROUGE-L

The figure below is a scattered chart between SARI and ROUGE-L for  $\beta = 0.1$ . It can be seen from the graph below that majority of the scores are centered around the mean values of SARI (38.41) and ROUGE-L (57.55), despite having several outliers in the cluster. The mean values of SARI and ROUGE-L scores are mentioned in Table 4.6.

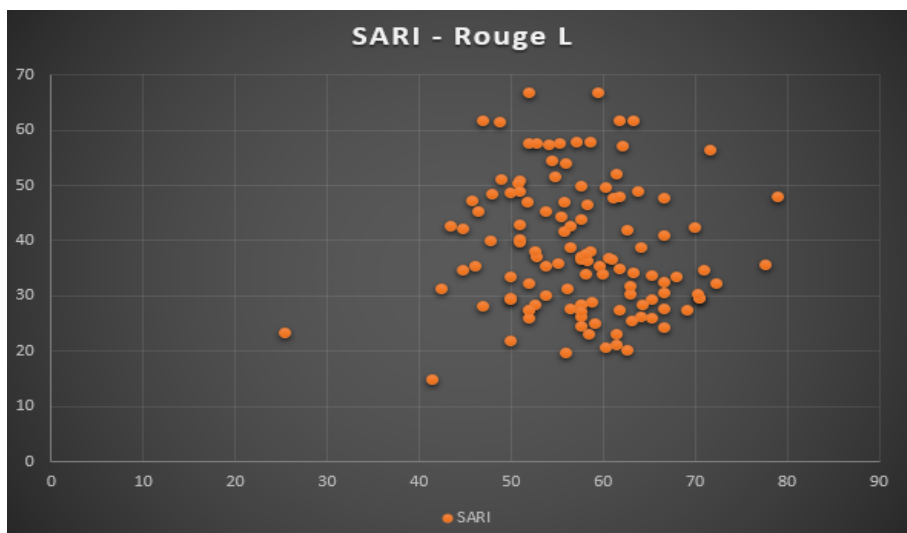


Figure 4.11: SARI, in relation with ROUGE L for  $\beta = 0.1$

### 4.2.3 For $\beta = 0.2$ :

#### SARI vs ROUGE-1

The figure below is a scattered chart between SARI and ROUGE-1 for  $\beta = 0.2$ . It can be seen from the graph below that majority of the scores are centered around the mean values of SARI (23.38) and ROUGE-1 (42.2), despite having several outliers in the cluster. The mean values of SARI and ROUGE-1 scores are mentioned in Table 4.6.

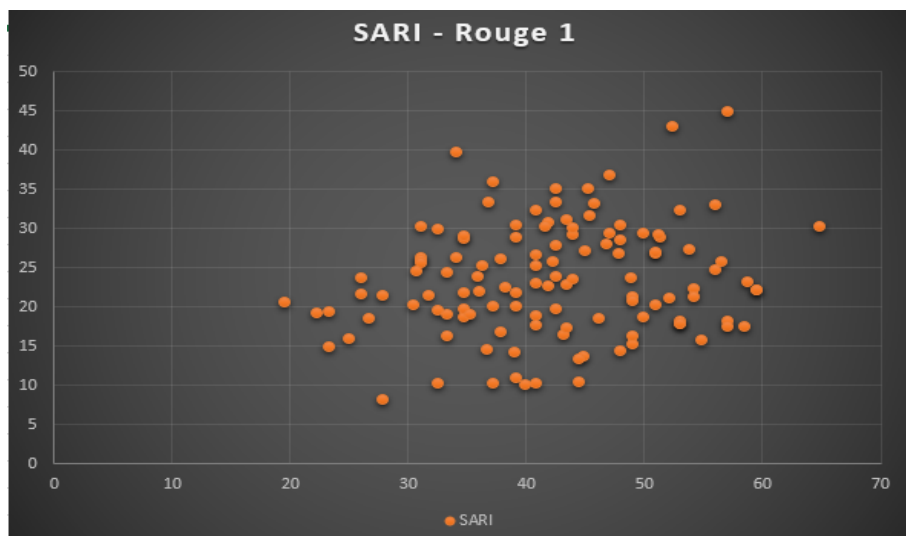


Figure 4.12: SARI, in relation with ROUGE I for  $\beta = 0.2$

#### SARI vs ROUGE-2

The figure below is a scattered chart between SARI and ROUGE-2 for  $\beta = 0.2$ . It can be seen from the graph below that majority of the scores are centered around the mean values of SARI (23.38) and ROUGE-2 (16.36), despite having several outliers in the cluster. The mean values of SARI and ROUGE-2 scores are mentioned in Table 4.6.

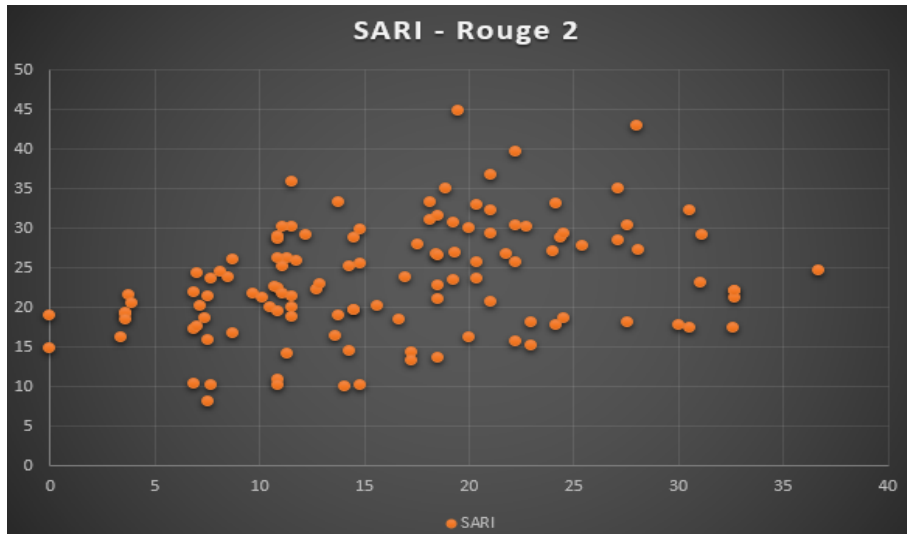


Figure 4.13: SARI, in relation with ROUGE II for  $\beta = 0.2$

### SARI vs ROUGE-L

The figure below is a scattered chart between SARI and ROUGE-L for  $\beta = 0.2$ . It can be seen from the graph below that majority of the scores are centered around the mean values of SARI (23.38) and ROUGE-L (41.71), despite having several outliers in the cluster. The mean values of SARI and ROUGE-L scores are mentioned in Table 4.6.

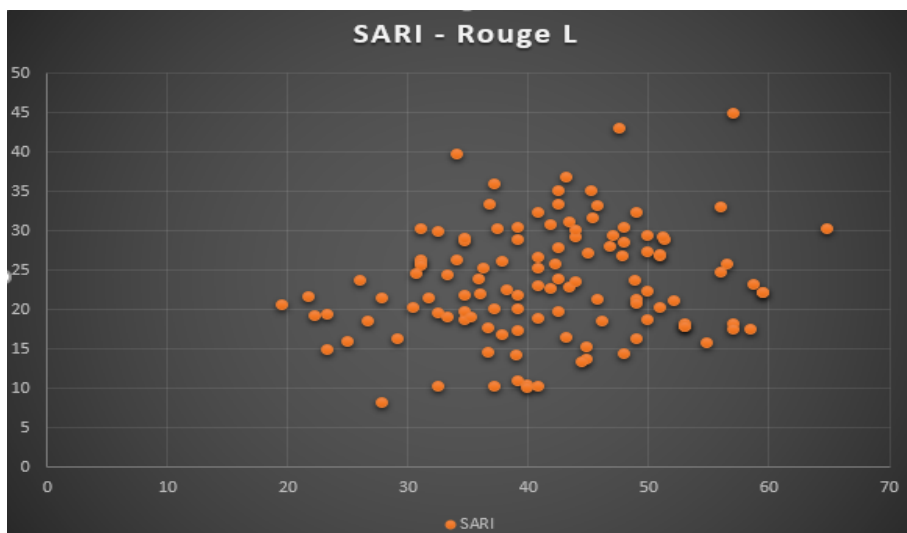


Figure 4.14: SARI, in relation with ROUGE L for  $\beta = 0.2$

#### 4.2.4 For $\beta = 0.3$ :

##### SARI vs ROUGE-1

The figure below is a scattered chart between SARI and ROUGE-1 for  $\beta = 0.3$ . It can be seen from the graph below that majority of the scores are centered around the mean values of SARI (32.77) and ROUGE-1 (52.47), despite having several outliers in the cluster. The mean values of SARI and ROUGE-1 scores are mentioned in Table 4.6.

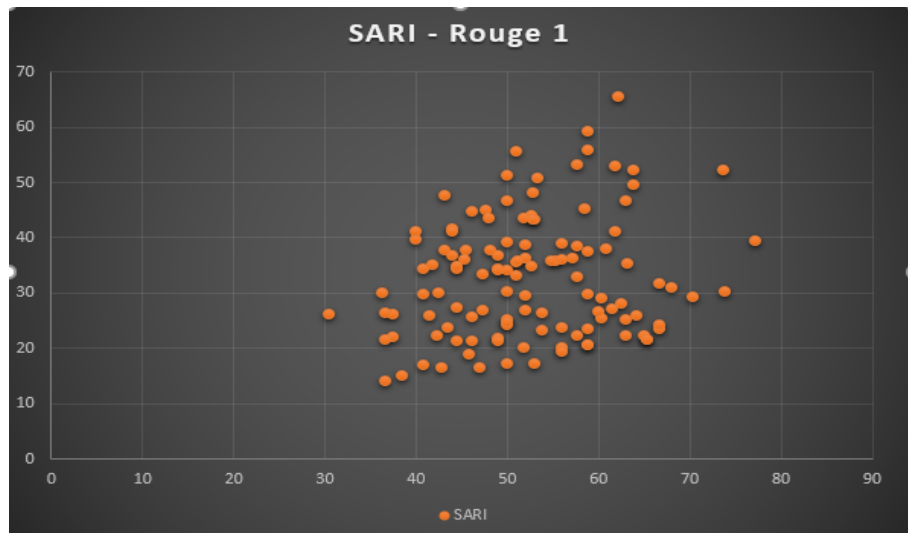


Figure 4.15: SARI, in relation with ROUGE I for  $\beta = 0.3$

##### SARI vs ROUGE-2

The figure below is a scattered chart between SARI and ROUGE-2 for  $\beta = 0.3$ . It can be seen from the graph below that majority of the scores are centered around the mean values of SARI (32.77) and ROUGE-2 (29.96), despite having several outliers in the cluster. The mean values of SARI and ROUGE-2 scores are mentioned in Table 4.6.

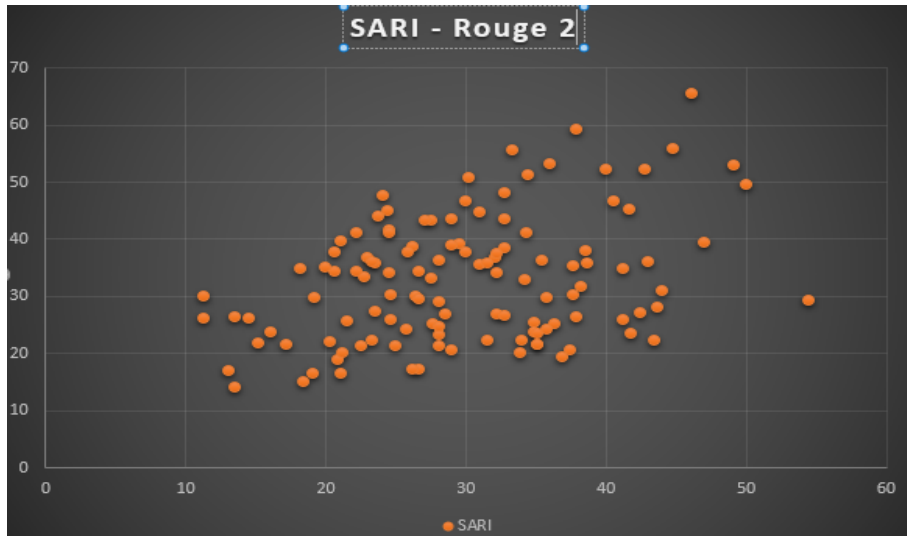


Figure 4.16: SARI, in relation with ROUGE II for  $\beta = 0.3$

### SARI vs ROUGE-L

The figure below is a scattered chart between SARI and ROUGE-L for  $\beta = 0.3$ . It can be seen from the graph below that majority of the scores are centered around the mean values of SARI (32.77) and ROUGE-L (52.25), despite having several outliers in the cluster. The mean values of SARI and ROUGE-L scores are mentioned in Table 4.6.

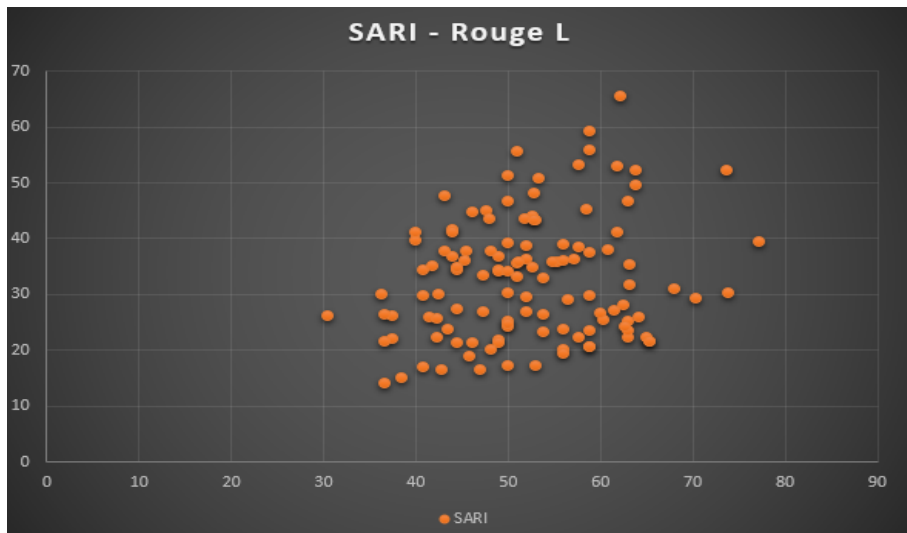


Figure 4.17: SARI, in relation with ROUGE L for  $\beta = 0.3$

### 4.2.5 For $\beta = 0.5$ :

#### SARI vs ROUGE-1

The figure below is a scattered chart between SARI and ROUGE-1 for  $\beta = 0.5$ . It can be seen from the graph below that majority of the scores are centered around the mean values of SARI (34.85) and ROUGE-1 (56.18), despite having several outliers in the cluster. The mean values of SARI and ROUGE-1 scores are mentioned in Table 4.6.

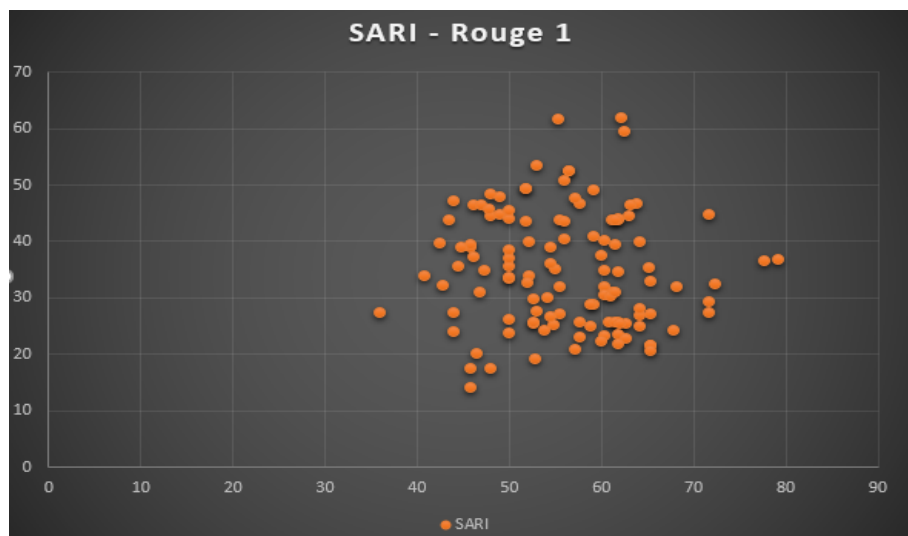


Figure 4.18: SARI, in relation with ROUGE I for  $\beta = 0.5$

#### SARI vs ROUGE-2

The figure below is a scattered chart between SARI and ROUGE-2 for  $\beta = 0.5$ . It can be seen from the graph below that majority of the scores are centered around the mean values of SARI (34.85) and ROUGE-2 (33.64), despite having several outliers in the cluster. The mean values of SARI and ROUGE-2 scores are mentioned in Table 4.6.

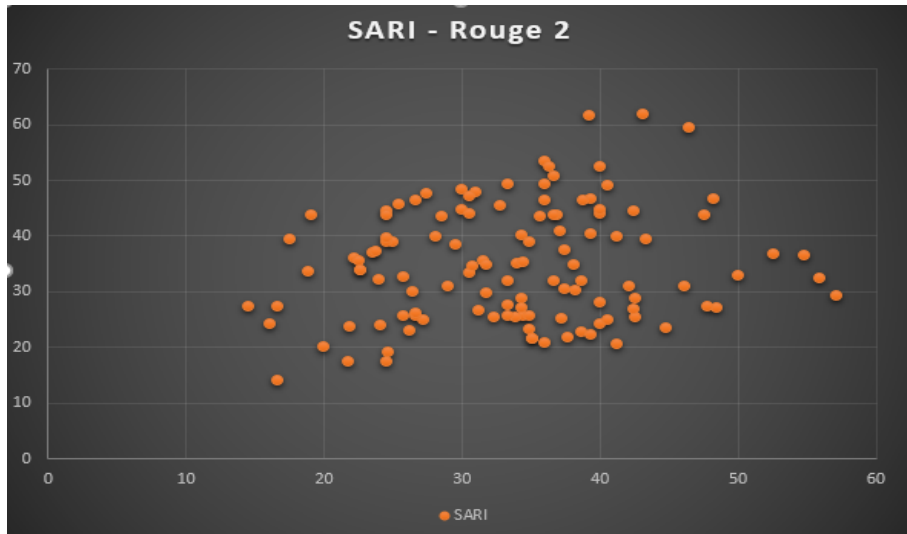


Figure 4.19: SARI, in relation with ROUGE II for  $\beta = 0.5$

### SARI vs ROUGE-L

The figure below is a scattered chart between SARI and ROUGE-L for  $\beta = 0.5$ . It can be seen from the graph below that majority of the scores are centered around the mean values of SARI (34.85) and ROUGE-L (55.84), despite having several outliers in the cluster. The mean values of SARI and ROUGE-L scores are mentioned in Table 4.6.

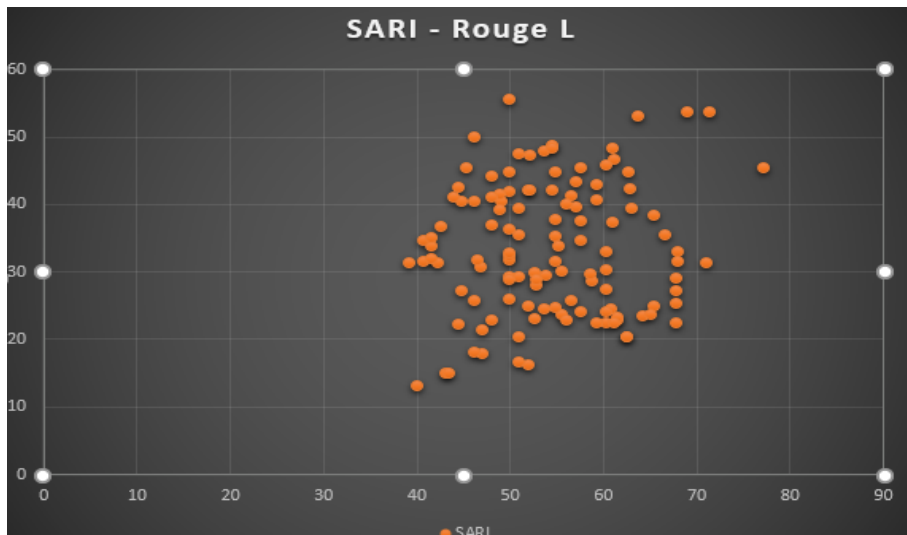


Figure 4.20: SARI, in relation with ROUGE L for  $\beta = 0.5$

### 4.2.6 For $\beta = 0.9$ :

#### SARI vs ROUGE-1

The figure below is a scattered chart between SARI and ROUGE-1 for  $\beta = 0.9$ . It can be seen from the graph below that majority of the scores are centered around the mean values of SARI (33.36) and ROUGE-1 (54.56), despite having several outliers in the cluster. The mean values of SARI and ROUGE-1 scores are mentioned in Table 4.6.

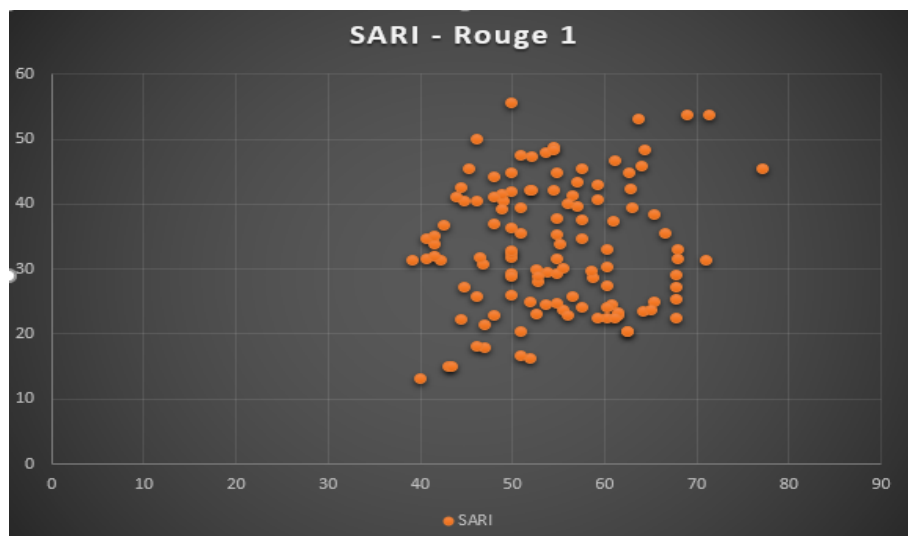


Figure 4.21: SARI, in relation with ROUGE I for  $\beta = 0.9$

#### SARI vs ROUGE-2

The figure below is a scattered chart between SARI and ROUGE-2 for  $\beta = 0.9$ . It can be seen from the graph below that majority of the scores are centered around the mean values of SARI (33.36) and ROUGE-2 (31.43), despite having several outliers in the cluster. The mean values of SARI and ROUGE-2 scores are mentioned in Table 4.6.



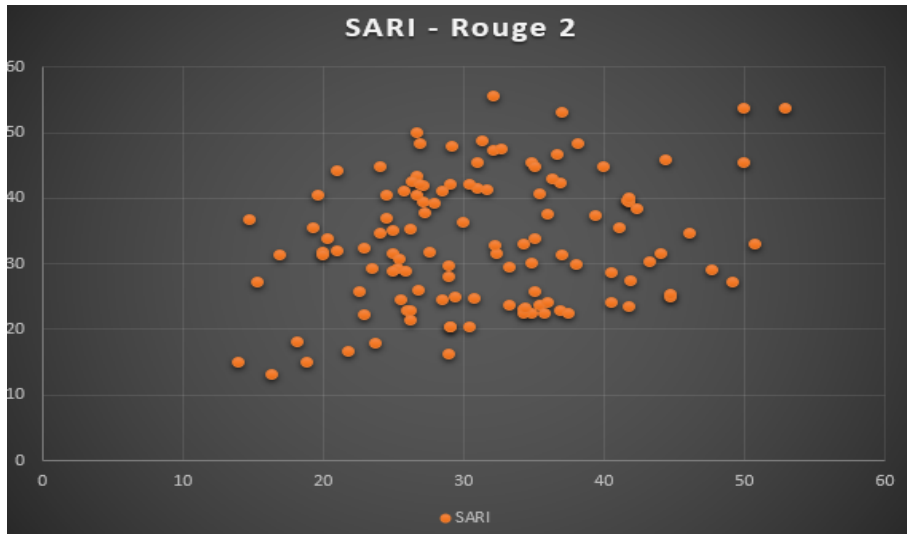


Figure 4.22: SARI, in relation with ROUGE II for  $\beta = 0.9$

### SARI vs ROUGE-L

The figure below is a scattered chart between SARI and ROUGE-L for  $\beta = 0.9$ . It can be seen from the graph below that majority of the scores are centered around the mean values of SARI (33.36) and ROUGE-L (54.47), despite having several outliers in the cluster. The mean values of SARI and ROUGE-L scores are mentioned in Table 4.6.

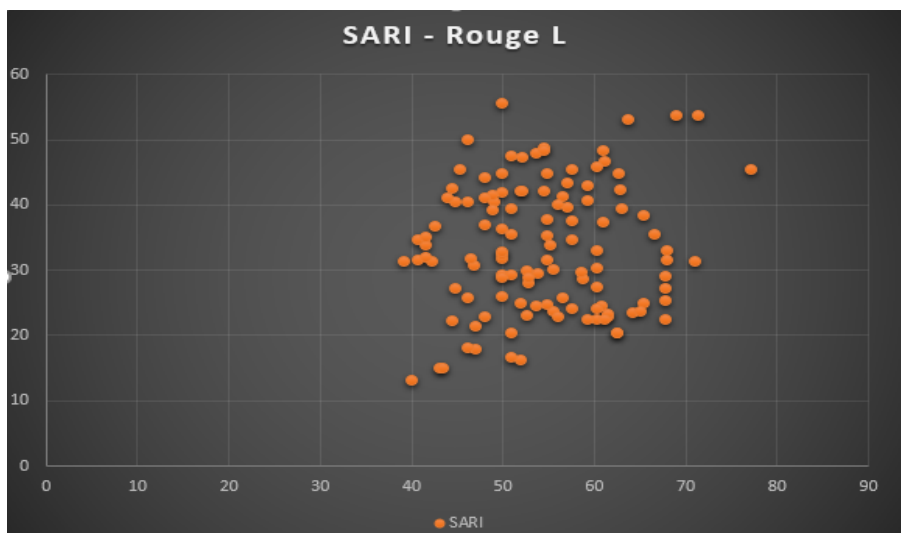


Figure 4.23: SARI, in relation with ROUGE L for  $\beta = 0.9$

### 4.3 Aggregated Evaluation

Finally, we evaluate the output of our summaries using SARI, ROUGE and CSS1 with respect to different values of  $\beta$ .

The table 4.6 shows mean values of SARI, ROUGE and CSS1 with respect to different values of  $\beta$ . This table also describes the calculated values of CSS-1 with respect to ROUGE-1 and SARI scores.

EVALUATION & SCORES TABLE											
$\beta$	ROUGE I			ROUGE II			ROUGE L			SARI	CSS I
	PRECISION	RECALL	FI-Score	PRECISION	RECALL	FI-Score	PRECISION	RECALL	FI-Score		
0.0	84.1	39.42	53.33	39.27	23.24	28.97	83.79	39.28	53.14	31.24	39.40
0.1	84.42	44.3	57.73	46.58	30.16	36.35	84.16	44.29	57.55	38.41	46.13
0.2	81.81	28.89	42.2	25.26	12.22	16.36	80.21	28.55	41.71	23.38	30.09
0.3	81.53	39.05	52.47	40.31	24.11	29.96	81.22	38.88	52.25	32.77	40.34
0.5	84.38	42.44	56.18	44.28	27.37	33.64	84.29	42.17	55.84	34.85	43.02
0.9	81.33	41.52	54.56	41.19	25.72	31.43	81.19	41.45	54.47	33.36	41.40

Table 4.6: SARI, ROUGE, and CSS1 scores in relation to  $\beta$

Diagram 4.6 shows the Rouge-1, SARI and CSS1 scores in relation to  $\beta$ . From this diagram, it can also be seen that all of the evaluation metrics performed well for  $\beta = 0.1$ , and performed worst for  $\beta = 0.2$ .

This suggests that the optimal value for our hyperparameter  $\beta$  is 0.1.

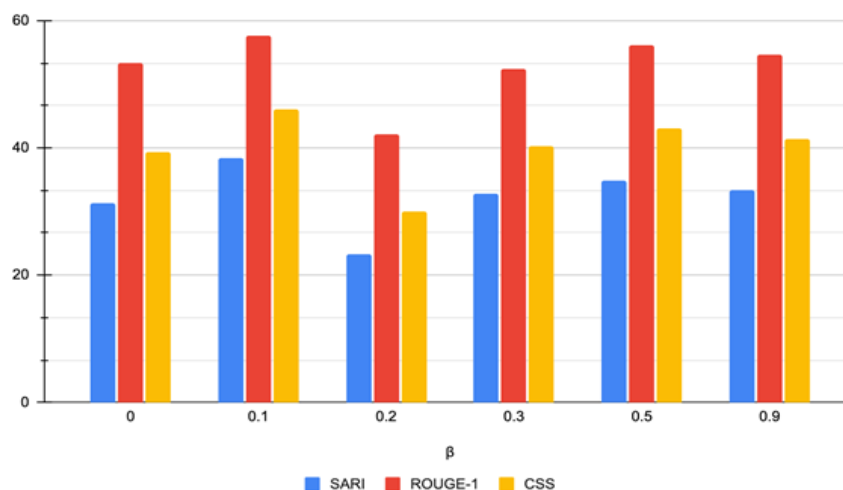


Figure 4.24: Rouge-1, SARI and CSS1 scores in relation to  $\beta$

System evaluation also suggested that our system performed better on ROUGE-1 and ROUGE-L, in comparison to ROUGE-2. ROUGE-2 calculates the commonality of bigrams between the source and target text, and is considerably lower than the other 2 metrics in our evaluation. This reveals that this was overall a harder task to simplify a summarized text while keeping the core information preserved.

## 4.4 Human Evaluation:

After getting the generated summaries evaluated by the system, we did human evaluation to verify the accuracy and correctness of the output of our system. Despite having good ROUGE, SARI and CSS scores, we came to know that the generated summaries were inaccurate and had a lot of room for improvement. Below is a table showing few of the summaries generated by our system.

Original-Summary	Generated-Summary
[ islamabad high court ] before aamer farooq , j messrs aimnaz ( pvt . ) limited versus federation of pakistan , through the secretary , ministry of law , federal secretariat , islamabad and 2 others writ petition no.2547	[UNK] islamabad [UNK] [UNK] [UNK] j before aamer j messrs aimnaz ( pvt pvt . ) versus federation of pakistan , , through the , , , , , , [UNK] [UNK] [UNK] [UNK] [UNK] [UNK] [UNK] [UNK]
[ islamabad ] before athar minallah , j telecom services and consultants ( pvt . ) ltd. -applicant versus ooredoo q.s.c . and others -respondents c.s .	[UNK] islamabad [UNK] before athar minallah j telecom services and consultants ( pvt . ) -- -- -- versus ooredoo q.s.c -- -- -- -- -- [UNK] [UNK] [UNK] [UNK] [UNK] [UNK] [UNK] [UNK] [UNK] [UNK] [UNK]
[ islamabad ] before athar minallah , j national feeds limited petitioner versus competition commission of pakistan and others respondents writ petition no . 1987 of 2015 , decided on 29th april , 2016 . ( a ) competition act	[UNK] islamabad [UNK] before athar minallah , j national feeds limited -- -petitioner versus competition commission of pakistan and others [UNK] others [UNK] [UNK] [UNK] [UNK] [UNK] [UNK] [UNK] [UNK] [UNK] [UNK] [UNK]
[ islamabad high court ] before aamer farooq and athar minallah , jj chief security officer airports security force and others versus tariq ahmed lodhi intra-court appeal no.1089 of 2013 , heard on 12th february , 2015 . ( a	[UNK] islamabad [UNK] [UNK] j before farooq athar minallah , jj chief security officer airports security force and others [UNK] tariq ahmed lodhi lodhi [UNK] [UNK] [UNK] [UNK] [UNK] [UNK] [UNK] [UNK]
[ islamabad high court ] before aamer farooq , j iftikhar rashid and 3 others versus federation of pakistan and 5 others writ petition no.1233 of 2015 , decided on 29th may , 2015 . constitution of pakistan -arts .	[UNK] islamabad [UNK] before before aamer farooq , j iftikhar and 3 others [UNK] federation pakistan and 5 5 others [UNK] others [UNK] [UNK] [UNK] [UNK] [UNK] [UNK] [UNK] [UNK] [UNK] [UNK] [UNK]

Figure 4.25: Human-written Summaries vs System-generated Summaries

The [UNK] tokens in the output summary tell that the system failed to find any suitable word for the given word. One possible reason for so many unknown tokens could be the fact that we had a very small vocabulary as

## CHAPTER 4. RESULTS AND DISCUSSION

compared to the original model that had a larger vocabulary. Another possible reason could be the fact that our dataset needed further preprocessing and proper classification.

# Chapter 5

## Conclusion and Future Work

*This chapter provides the conclusion and future work of the thesis.*

### 5.1 Conclusion

In this research, we applied HTSS algorithm on the dataset consisting the hearings of Supreme Court of Pakistan. The dataset contained the court hearings as input source and summaries as the ground truth. Our task was to generate simplified summaries using the HTSS algorithm. We trained our model on various values of hyperparameter  $\beta$  and concluded that overall system gave the best performance when  $\beta = 0.1$ .

The system evaluation suggested that our system performed better on ROUGE-1 and ROUGE-L, in comparison to ROUGE-2. ROUGE-2 calculates the commonality of bigrams between source and target text and is considerably lower than the other 2 metrics in our evaluation, showing that this was a harder task all round.

However, the human evaluation revealed that the generated summaries were not very accurate and required more improvement.

### 5.2 Future Work

In future, we could further preprocess the dataset and annotate it in a similar way as Zaman et al. did in their original implementation. Another future

## *CHAPTER 5. CONCLUSION AND FUTURE WORK*

prospect of this research is to build vocabulary and weighted words for legal cases and see the results. Another approach would be to use transfer learning, learn from a legal database model on legal dataset and apply that trained model on our dataset. Also, we will pursue to improve the loss function by integrating further characteristics that indicate the complexity of a word for example length, frequency, and concentration. We will also ponder how to increase the vocabulary of our model to help it to prevent inserting ‘UNK’ tokens into the productivity.

# Bibliography

- [1] H. Oliveira, R. D. Lins, R. Lima, F. Freitas, and S. J. Simske, “A regression-based approach using integer linear programming for single-document summarization,” *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, vol. 2017-November, pp. 270–277, 2018.
- [2] Y. Zhang, J. Liao, J. Tang, W. Xiao, and Y. Wang, “Extractive document summarization based on hierarchical GRU,” *Proceedings - 2018 International Conference on Robots and Intelligent System, ICRIS 2018*, pp. 341–346, 2018.
- [3] —, “Extractive document summarization based on hierarchical GRU,” *Proceedings - 2018 International Conference on Robots and Intelligent System, ICRIS 2018*, pp. 341–346, 2018.
- [4] J. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Pérez, “An Indicator-based Multi-Objective Optimization Approach Applied to Extractive Multi-Document Text Summarization,” *IEEE Latin America Transactions*, vol. 17, no. 8, pp. 1291–1299, 2019.
- [5] S. Ghodratnama, A. Beheshti, M. Zakershaharak, and F. Sobhanmanesh, “Extractive Document Summarization Based on Dynamic Feature Space Mapping,” *IEEE Access*, vol. 8, pp. 139 084–139 095, 2020.
- [6] Rahul, S. Adhikari, and Monika, “Nlp based machine learning approaches for text summarization,” in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 2020, pp. 535–538.

## BIBLIOGRAPHY

- [7] F. L. G. Ayesha Ayub Syed and T. Matsu, “A survey of the state-of-the-art models in neural abstractive text summarization,” *IEEE Access*, p. 13248–13265, 2021.
- [8] M. Yang, C. Li, Y. Shen, Q. Wu, and Z. Zhao, “Hierarchical human-like deep neural networks for abstractive text summarization,” *IEEE Transaction Neural network*, p. 2744–2757, 2021.
- [9] F. Z. Jingwei Cheng and X. Guo, “A syntax-augmented and headline-aware neural text summarization method,” *IEEE Access*, p. . 218360–218371, 2020.
- [10] K. S. Vu D. Tran, Minh L. Nguyen and K. Satoh, “An approach of rhetorical status recognition for judgments in court documents using deep learning models,” *11th International Conference on Knowledge and Systems Engineering (KSE) - IEEE*, 2019.
- [11] K. Kowsrihawatt and P. Vateekul, “An information extraction framework for legal documents: a case study of thai supreme court verdicts,” *12th International Joint Conference on Computer Science and Software Engineering (JCSSE) - IEEE*, pp. 275–280, 2015.
- [12] Y. Ma, P. Zhang, , and J. Ma, “An ontology driven knowledge block summarization approach for chinese judgment document classification,” *Data Mining and Granular Computing in Big Data and Knowledge Processing = IEEE*, pp. 71 327 – 71 338, 2018.
- [13] M. D. B. Deepali Jain and A. Biswas, “Automatic summarization of legal bills: A comparative analysis of classical extractive approaches,” *International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) - IEEE*, pp. 394–400, 2021.
- [14] D. H. Chandrika Prasad, Jagdish S. Kallimani and N. Sharma, “Automatic text summarization model using seq2seq technique,” *Proceedings of the Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)-IEEE*, pp. 599–604, 2020.



## BIBLIOGRAPHY

- [15] S. T. Hiroaki Yamada and T. Tokunag, “Designing an annotation scheme for summarizing japanese judgment documents,” *9th International Conference on Knowledge and Systems Engineering(KSE)*, pp. 275–280, 2017.
- [16] A. Trivedi, A. Trivedi, S. Varshney, V. Joshipura, R. Mehta, and J. Dhanani, “Extracted summary based recommendation system for indian legal documents,” *International Conference on Computing, Communication and Networking Technologies (ICCCNT)-IEEE*, 2020.
- [17] K. Merchant and Y. Pande, “Nlp based latent semantic analysis for legal text summarization,” *IEEE Access*, pp. 1803–1807, 2021.
- [18] F. Zaman, M. Shardlow, S.-U. Hassan, N. R. Aljohani, and R. Nawaz, “Htss: A novel hybrid text summarisation and simplification architecture,” *Information Processing & Management*, vol. 57, no. 6, p. 102351, 2020.
- [19] P. J. L. Abigail See and C. D. Manning, “Get to the point: Summarization with pointer-generator networks,” *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, vol. 1, p. 1073–1083, 2017.
- [20] N. I. Nikolov, M. Pfeiffer, and R. H. Hahnloser, “Data-driven summarization of scientific articles,” *arXiv preprint arXiv:1804.08875*, 2018.
- [21] R. Nallapati, B. Zhou, C. N. dos santos, C. Gulcehre, and B. Xiang, “Abstractive text summarization using sequence-to-sequence rnns and beyond,” 2016.
- [22] O. Vinyals, M. Fortunato, and N. Jaitly, “Pointer networks,” in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: <https://proceedings.neurips.cc/paper/2015/file/29921001f2f04bd3baee84a12e98098f-Paper.pdf>

## BIBLIOGRAPHY

- [23] S. Nisioi, S. Štajner, S. P. Ponzetto, and L. P. Dinu, “Exploring neural text simplification models,” in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)*, 2017, pp. 85–91.