ENHANCING THE EXPLAINABILITY OF TRANSFORMER-BASED ABSTRACTIVE SUMMARIZATION MODELS

P. H. Panawenna

239167T

Master of Science in Data Science and Artificial Intelligence

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

May 2025

ENHANCING THE EXPLAINABILITY OF TRANSFORMER-BASED ABSTRACTIVE SUMMARIZATION MODELS

P. H. Panawenna

239167T

Dissertation submitted in partial fulfillment of the requirements for the degree
Master of Science in Data Science and Artificial Intelligence

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

May 2025

DECLARATION

I declare that this is my own work and this Dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 02-07-2025

The supervisor should certify the Dissertation with the following declaration.

The above candidate has carried out research for the Master of Science in Data Science and Artificial Intelligence Dissertation under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr. Sandareka Wickramanayake

Signature of the Supervisor: Date: 04/07/2025

DEDICATION

This report is dedicated to my parents for their unwavering support and unconditional love throughout the years.

ACKNOWLEDGEMENT

First and foremost, I would like to extend my heartfelt gratitude to my supervisor, Dr. Sandareka Wickramanayake, for her invaluable insights, unwavering support, and motivating guidance throughout this research. Her dedication inspired me, as she supported me during countless late nights. She was available whenever I had questions, and constantly encouraged me to push my limits. The opportunities she provided for publication and her mentorship in every aspect of the project have shaped me into a better researcher.

I would also like to sincerely thank Prof. Dulani Meedeniya for her valuable insights and continued support from the very beginning. Her guidance has been an important part of this research, along with the opportunities she provided for publications.

My heartfelt appreciation goes to Mr. Kasun Gayashan Hettihewa, who worked as a Research Assistant at the Department of Computer Science and Engineering. He contributed as a co-author in publications related to this research. His generous and consistent support has been instrumental in navigating the challenges of this project. His work in comparing the effectiveness of different feature attribution methods in explaining transformer-based abstraction summarization, presented as one of the sections in our publication [1], sheds light on the utility of the explanation framework introduced in this research.

To my parents, thank you for being my constant pillars of strength, standing by me through both the highs and lows of this journey.

To my friends and family, your patience, understanding, and words of encouragement were crucial in helping me balance the demands of a full-time job while pursuing my MSc Degree. I am deeply grateful for your support.

I would also like to acknowledge my workplace, ZeroBeta (Pvt) Ltd., for funding my MSc degree and for providing the necessary study leave, which enabled me to dedicate time and effort to this research.

Finally, I am truly thankful to the medical professionals who participated in the user study of this research. Despite their demanding schedules, they took the time to offer their honest opinions and expert insights. Their contributions added immense value to the evaluation and validation of this work.

To all who supported me along the way, thank you.

ABSTRACT

Abstractive Summarization (AS) is a Natural Language Processing (NLP) task that generates a concise and coherent summary of a given document by rephrasing or paraphrasing the content. It captures the essential information rather than directly extracting sentences or phrases from the source text, as opposed to Extractive Summarization (ES). AS is used in multiple mission-critical domains such as healthcare, law, and finance. Nevertheless, the existing state-of-the-art AS models are based on black-box deep learning models such as Transformers, and they cannot explain why specific facts were included in the summary while some facts were omitted. This research proposes a novel framework to explain which facts have been excluded from the summary by a given AS model and the rationale behind the selections. The new framework, Fact Omission Explanation (FOE) , utilizes a feature attribution method to analyze the fact-selection process of a given AS model and generate a linguistic explanation of which facts have been excluded and the respective reasons. The proposed framework was assessed using the PubMed dataset and Arxiv dataset, which consists of long documents in medical and scientific domains, and PEGASUS and T5 transformer models, which are state-of-the-art transformer-based AS models. A user study was conducted with the participation of medical professionals to assess the value addition of the framework in practice. The results demonstrate that the generated explanations help ensure the trustworthiness of AS models in mission-critical domains such as healthcare.

Keywords: Abstractive Summarization, Natural Language Processing, Transformers, Explainable AI

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
AI	Artificial Intelligence
AMR	Abstract Meaning Representation
AS	Abstractive Summarization
BART	Bidirectional and Auto-Regressive Transformers
BERT	Bidirectional Encoder Representations from Transformers
CA	Cross Attention
CNN	Convolutional Neural Networks
DL	Deep Learning
DNN	Deep Neural Networks
ERASER	Evaluating Rationales And Simple English Reasoning
ES	Extractive Summarization
FOE	Fact Omission Explanation
GI	Gradient * Input
GPT	Generative Pre-Trained Transformers
Grad-CAM	Gradient-weighted Class Activation Mapping
GRU	Gated Recurrent Unit
GSG	Gap Sentences Generation
GSR	Gap Sentences Ratio
ILP	Integer Linear Programming
INITs	INformation ITems
LED	Longformer Encoder Decoder
ML	Machine Learning
MLM	Masked Language Model
NLP	Natural Language Processing
RNN	Recurrent Neural Networks
SA	Self Attention
SP	Summarization Programs
T5	Text-to-Text Transfer Transformers
XAI	Explainable Artificial Intelligence