# TEXT MINING (NLP)
# ABSTRACTIVE TEXT SUMMARIZATION USING DEEP SEQUENCE MODELS

**ENGR. MUHAMMAD IRFAN**

**01-242182-005**

A thesis submitted in the fulfillment of the Requirements for the award of the degree of Master of Science (Computer Engineering)

**Department of Computer Engineering**

**BAHRIA UNIVERSITY ISLAMABAD**

**2021**

## APPROVAL FOR EXAMINATION

Scholar's Name: Engr. Muhammad Irfan Registration No. 01-242172-006

Programmed of Study: MSCE

Thesis Title: Abstractive Text Summarization Using Deep Sequence Models

It is to certify that the above thesis has been completed to my satisfaction and my belief. Its standard is appropriate for exam submission. I have also conducted a plagiarism test of this thesis using HEC prescribed software and found a similarity index of about 17% that is within the permissible limit set by the HEC for the MS degree thesis. I have also found the thesis in a format recognized by the BU for the MS thesis.

**Principal Supervisor's Signature: ___________________________________**

**Date: _________________**

**Name: ___________________________________**

## DECLARATION

I, Muhammad Irfan solemnly states that my MS thesis Abstractive Text Summarization Using Deep Sequence Models is my effort. It has no such material which is earlier published. All the mentions and necessary help in this study have been recognized. I affirm that the material in this research has not been used by me from Bahria University for another degree at any other institution.

Name of the scholar: Engr. Muhammad Irfan

Date:______________

# PLAGIARISM UNDERTAKING

I, Muhammad Irfan solemnly declare that research work presented in the thesis titled "Abstractive Text Summarization Using Deep Sequence Models" is solemnly my research work with no significant contribution from any other person. Small contribution helps wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero-tolerance policy of the HEC and Bahria University towards plagiarism. Therefore, I as an author of the above-titled thesis declare that no portion of my thesis has been plagiarized and any material used as the reference is properly referred /cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled even after awarding of MS degree. The university reserves the right to withdraw /revoke my MS degree and HEC and the University has the right to publish my name on the HEC/University website on which names of scholars are placed who submitted plagiarized thesis.

Author's Signature

Name of the Scholar: Engr. Muhammad Irfan

# DEDICATION

To My Father, Mother, all my Family, and Teachers.

# ACKNOWLEDGMENT

First of all I would like to thank Prof. Dr. Shehzad Khalid (Director ORIC BU Islamabad) and Dr. Muhammad Taimoor Khan (Assistant Professor FAST - NUCES Peshawar) for giving me the opportunity to carry out state-of-the-art research in this field. Furthermore, I would like to thank for supervision of my thesis. Finally, I must express my very profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

# Abstract

The reading of the long text is time consuming and sometimes understanding the context becomes difficult. Summaries are important specifically when we need to save our time and to understand the actual context of a long text corpus. Summarization is a technique to create a concise and accurate summary of a large script or a set of articles. In recent years abstractive text summarization tasks are most challenging in natural language processing. The existing encoder-decoder approaches have a potential issue. For the longer sequence of reviews, they need to compress all the necessary information into a fixed-length vector. This thesis aims to solve a very inherent task in data mining that is review summarization. Summary of the reviews has challenges that are dealing with variable length reviews, free-style writing, and unstructured behavior. Our aim to create a shorter version of the review in abstractive manners while preserving the sentiment and points. In the decision-making process, it helps online customers to judge the product or service. To generate an optimal summary we have used a BRNN with LSTM's in the encoding layer. In the decoding layer, the attention mechanism is applied to the decoding cell that is just a two-layer LSTM with dropout. We have used ConceptNet Number-Batch 3.0 word embeddings and Amazon Food reviews dataset. To reduce training loss and compute the learning rate of each parameter, we have used Adam Optimizer to reduce the loss function and for faster converges. We have achieved R1 38.75, R2 16.5, RL 36.25, and reduced the training loss with a new value of **0.031** for the whole dataset after removing the duplications.

**Keywords:** *Abstractive Text Summarization, BRNNs, LSTMs, Attention Model.*

# Table of Contents

**Contents**

Title Page ........ i
Approval for Examination ........ ii
Declaration ........ iii
Plagiarism Undertaking ........ iv
Dedication ........ v
Acknowledgment ........ vi
Abstract ........ vii
Table of Contents ........ viii
List of Figures ........ x
List of Tables ........ xi
List of Symbols and Abbreviations ........ xii

1 Introduction 1
- 1.1 Problem Statement ........ 3
- 1.2 Overview of our Proposed Methodology ........ 3
- 1.3 Organization of the Thesis ........ 4

2 Literature Review 5
- 2.1 Fundamental Models Employed for Abstractive Text Summarization ........ 11
- 2.1.1 RNNs Sequence Models ........ 12
- 2.1.2 Bidirectional RNNs ........ 16
- 2.1.3 Long Short Term Memory Networks (LSTM) ........ 16
- 2.1.4 Word Embeddings ........ 19
- 2.1.5 Sequence to Sequence Neural Networks ........ 20

3 Methodology 24
- 3.1 Objective ........ 25
- 3.2 Proposed Idea ........ 25
- 3.3 Proposed Model ........ 25

3.4 Various Phases in Developing Abstractive Review Summarization ....28
3.4.1 Inspection and Pre-Processing ....29
3.4.2 Selecting Words Embedding ....29
3.4.3 Basic Encoder-Decoder ....30
3.4.4 Attention Model ....31
3.4.5 The Encoding Layer Building ....32
3.4.6 The Decoding Layer Building ....32
3.4.7 Encoder's General Description ....33
3.4.8 Decoder's General Description ....33
3.4.9 Attention Mechanism Intuition ....35
3.4.10 Developing Abstractive Review Summary ....36
4 Experiments and Results 37
Discussions 44
5 Conclusion and Future work 46
References 48

# List of Figures

Figure 2.1 Abstractive Summarization Approaches ........ 9

Figure 2.6 Bidirectional RNN ........ 16

Figure 2.8 LSTM Intuitive Illustration ........ 18

Figure 2.9 Seq2seq Model Architecture ........ 21

Figure 3.1 Proposed Model for Abstractive Text Summarization ........ 26

Figure 3.2 Various Phases of Abstractive Review Summarization ........ 28

Figure.3.3 Concept Net Number-Batch shows less Demographic Bias ........ 30

Figure 3.5 Encoder-Decoder General Descriptions ........ 35

Figure 4.1. Shows the Length of the Input Text and Output Summary for the Entire Dataset ........ 39

Figure .4.2. Graphical Representation of the Training Loss ........ 41

Figure .4.3. Analyzed Result Comparison ........ 43

# List of Tables

Table 2.1 An analytical table of the research work leads to the problem statement.......10

Table 4.1 Exploratory Data Analysis.......40

Table 4.2 Training Loss Result Comparison.......41

Table 4.3 Generated Summaries.......42

Table 4.4 Performance Metrics ROUGE, F1and BLEU Scores.......42

Table 4.5 Cosine Similarity Scores.......43

Table 4.6 Analyzed Result Comparison.......43

# List of Symbols and Abbreviations

| | |
|---|---|
| ANN | Artificial Neural Network |
| BLEU | Bilingual Evaluation Understudy |
| BRNN | Bi Directional Recurrent Neural Network |
| BLSTM | Bi Directional Long Short-Term Memory |
| BERT | Bidirectional Encoder Representations from Transformers |
| CN | ConceptNet |
| CNN | Convolutional Neural Network |
| DL | Deep Learning |
| GRU | Gated Recurrent Unit |
| GloVe | Global Vectors for Word Representation |
| HDF | High Definition File |
| IR | Information Retrieval |
| IDF | Inverse Document Frequency |
| LSTM | Long Short-Term Memory |
| NLP | Natural language Processing |
| NLTK | Natural Language Tool Kit |
| OOV | Out of Vocabulary |
| RE | Regular Expression |
| RL | Reinforcement Learning |
| RNN | Recurrent Neural Network |
| ROUGE | Recall-Oriented Understudy for Gisting Evaluation |
| SEO | Search Engine Optimization |
| Seq2seq | Sequence to Sequence |
| TF | Term Frequency |