Alzheimer's Disease using BERT Text Classification

Author

Shiza Latif

Registration Number: 000363860

Supervised by

Dr. Naeem Ul Islam

DEPARTMENT OF ELECTRICAL ENGINEERING

COLLEGE OF ELECTRICAL AND MECHANICALENGINEERING(E&ME)

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY(NUST)

ISLAMABAD

JULY,2023

THESIS ACCEPTANCE CERTIFICATE

It is certified that final copy of MS/MPhil thesis written by Ms. Shiza Latif (Registration No. 00000363860) Entry-2021, of (College of E&ME) has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors, and mistake and is accepted as partial fulfillment for award of MS degree. It is further certified that necessary amendments as pointed out by GEC member of the scholar have also been incorporated in the said thesis.

Signature: _____

Name of Supervisor Dr. Naeem Ul Islam

Date: 14-07-2023

Signature (HoD): _____

Dr. Fahad Mumtaz Malik

Date: 14-Jul-2023

Signature (Dean): _____

Brig Dr. Nasir Rashid

Date: 14 JUL 2023

Dedication

This success is a tribute to my extraordinary parents, ***Mr. & Mrs. Muhammad Latif****, whose constant support and collaboration made it possible for me to attain this achievement.*

Acknowledgements

All thanks and glory are due to Almighty Allah(the Most Gracious and Most Merciful), Who provided me with the strength, perseverance, knowledge, and skills necessary to carry out this task and endure until it was effectively completed. Undoubtedly, HE paved the route for me, and without HIS favor, I am powerless.

I would like to convey sincere appreciation to my adviser, Dr. Naeem ul Islam, for raising my spirits and for his ongoing support, inspiration, commitment, and priceless guidance in my pursuit of knowledge. I consider myself fortunate to have such a excellent mentor and a gracious advisor for my studies.

Along with my adviser, I would like to thank Dr. Mazhar Abbas and Dr. Salman Qadir from my thesis committee for their collaboration and thoughtful ideas.

Without appreciating my parents and my dear husband, my in-laws who are the main source of my strength, my appreciation would be lacking. I owe a great deal of gratitude to my devoted parents, who raised me from the time I was unable to walk and supported me in every aspect of my life, as well as to my caring sisters, who stood by me through good times and bad.

Finally, I want to thank all my friends and the people who have helped and supported me over this entire time.

Abstract

The progression of Alzheimer's disease is relentless, leading to a worsening of mental faculties over time. Currently, there is no remedy for this illness. Accurate detection and prompt intervention are pivotal in mitigating the progression of the disease. Recently, researchers have been developing new methods for detecting Alzheimer at earlier stages, including genetic testing, blood tests for biomarkers, and cognitive assessments. Cognitive assessments involve a series of tests to measure memory, language, attention, and other brain functions. Although there is still no definitive test for Alzheimer, research is ongoing and new techniques are being developed. For disease detection, optimal performance necessitates enhanced accuracy coupled with efficient computational capabilities. Our proposition involves, after data augmentation of textual data from Kaggle , it will then be analyzed using a BERT-based deep learning model in an effort to take use of its advanced capabilities for improved feature extraction and text comprehension. Our model is able to accurately detect Alzheimer's disease from textual data. We conduct a thorough assessment of our proposed BERT-based deep learning model for text categorization using a dataset made up of patient-reported medical records in order to determine its efficacy. In our comparison analysis, we compare our model to cutting-edge machine learning techniques frequently used for text classification tasks, such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and long short-term memory (LSTM) networks. This thorough evaluation intends to evaluate the performance and superiority of our suggested model in terms of precision, computational effectiveness, and capability to successfully capture complex textual patterns in the medical domain. Our result showed that our BERT-based model outperforms previously implemented methods based on BERT-CNN and BERT-RCNN in terms of accuracy, precision, and recall. Additionally, we used ensemble state-of-art techniques that allowed us to leverage the collective intelligence of the ensemble and make highly accurate and reliable predictions on the dataset.

Keywords—*Transformer, BERT, Alzheimer's Disease(AD), Deep Learning Model, Language and speech, NLP, Text classification*

Table of Contents

ACKNOWLEDGEMENTS ..... **V**

ABSTRACT ..... **VI**

LIST OF FIGURES ..... **IX**

LIST OF TABLES ..... **X**

CHAPTER 1 ..... **1**

INTRODUCTION **2**

1.1. OVERVIEW 2

1.2. PROBLEM STATEMENT 2

1.3. OBJECTIVES OF THE STUDY 3

1.4. THESIS ORGANIZATION 3

CHAPTER 2 ..... **4**

RISK FACTORS CONTRIBUTING TO ALZHEIMER'S DISEASE: A COMPREHENSIVE INVESTIGATION **5**

2.1 ALZHEIMER'S DISEASE (AD) AND ITS IMPACTS 5

2.2 SYMPTOMS 5

2.3 MEMORY LOSS 6

2.4 BEHAVIOR AND PERSONALITY CHANGES 6

2.5 CAUSES OF ALZHEIMER'S DISEASE 6

CHAPTER 3 ..... **8**

LITERATURE REVIEW **9**

3.1 OVERVIEW-BACKGROUND STUDY 9

3.2 AD DETECTION USING DEEP LEARNING 10

3.3 AD DETECTION USING TRANSFORMERS 12

3.4 DATA SET 13

CHAPTER 4 ..... **15**

4.1 PROPOSED METHODOLOGY 16

4.1.1 BERT *16*

4.1.2 What makes BERT efficient? *16*

4.1.3 Pre-process *18*

4.1.4 Encoder *20*

4.2 Optimizer *21*

4.3 Training Testing Split *22*

4.4 DATA AUGMENTATION 23

4.5 NEURAL NETWORK 24

4.6 RECURRENT NEURAL NETWORK: 26

4.7 RCNN 29

4.8 IMPLEMENTATION 31

4.9 ENSEMBLE 32

4.10 ALGORITHM 33

4.11 BLOCK DIAGRAM: 34

4.12 EXPERIMENT RESULTS: 34

4.13 ANALYZING GRAPH DATA FOR TECHNICAL INSIGHTS: 36

CHAPTER 5 39

5.1 DISCUSSION 40

CHAPTER 6 42

6.1 CONCLUSION AND FUTURE WORK 43

REFERENCES 45

LIST OF FIGURES

Figure 1: Structure of Brain 5
Figure 2: The histopathological signs of Alzheimer 7
Figure 3: Reference Image 14
Figure 4: Dataset analysis 14
Figure 5: Preprocess 19
Figure 6: Data Augmentation technique 24
Figure 7: RNN 27
Figure 8: Flow chart 33
Figure 9: Block diagram 34
Figure 10: CNN Accuracy 36
Figure 11: CNN Loss 36
Figure 12: RCNN Accuracy 37
Figure 13: RCNN Loss 37

LIST OF TABLES

Table 1: Concordance between predicted label and ground truths..... 30
Table 2: Results on Pitt Dataset..... 35
Table 3: Comparison of Classification Scores on Pitt datasets 41

CHAPTER 1

INTRODUCTION

1.1. Overview

Alzheimer is a degenerative illness that causes deterioration of the brain, resulting in memory loss and other cognitive impairments [1]. Studies have revealed that AD has a significant influence on the patient's language comprehension in addition to its effects on patients' mood, attention, memory, mobility, etc. Alzheimer is a progressive neurodegenerative disorder that affects recall, thinking pattern, and behavioral tendencies. It is the most common cause of dementia, accounting for 60-80% of all cases. The disease typically develops in people over the age of 65, and the risk of developing Alzheimer's increases with age. The Boston Aphasia Diagnostic Test's visual description task [2] has been demonstrated to be resistant to moderate cognitive deficits [3], suggesting that relevant medical records can be gleaned from voice sounds to recognize AD. Diagnosing AD using conversation transcripts has been shown to be feasible. It has been demonstrated that language analysis can be helpful in the evaluation of Vascular dementia, and despite utilizing a switcher to do features refinement and an analyzer to produce good results [4] , there is potential for improvement. In current research work, artificial intelligence has been developed as a promising tool to assist in the early diagnosis, treatment and prevention of AD. AI technology is especially effective when combined with deep learning algorithms, which make use of large amounts of training data.

1.2. Problem Statement

The existing methods for detecting Alzheimer's disease lack reliability and often involve invasive procedures, leading to a substantial number of incorrect diagnoses and missed opportunities for intervention. Additionally, these methods often rely on large datasets, further complicating the accurate identification of the disease. Improving the accuracy and non-invasive nature of Alzheimer's disease detection is therefore critical for enhancing patient outcomes and overall quality of life.

1.3. Objectives of the Study

The objective of this research is to develop a deep learning model based on BERT-based text classification for the correct detection of Alzheimer's using textual information. The suggested model intends to show its efficacy in terms of recall, accuracy, and precision. This study aims to demonstrate the model's potential as a powerful tool for the early identification of Alzheimer's disease by utilizing deep learning techniques and natural language processing approaches. The research will involve training the model on a diverse dataset of textual data associated with Alzheimer's disease and evaluating its performance using appropriate evaluation metrics. The ultimate objective is to give medical professionals a trustworthy, non-invasive way to reliably identify Alzheimer's disease at an early stage, enabling prompt therapies and better patient outcomes.

1.4. Thesis Organization

This work is structured as follows:

Chapter 2 describes symptoms of Alzheimer's disease in humans. It also goes into further detail on how and where it impacts the human body.

Chapter 3 discusses a review of the literature and major research into the early diagnosis of Alzheimer's disease is provided.

Chapter 4 gives the proposed methodology is described in depth. It has two basic modules: first, it segments the lungs, then it classifies them.

Chapter 5 introduces the models that will be utilized for evaluation. With all required figures and tables, the full discussion of all the experimental findings is provided.

Chapter 6 wraps up the thesis and describes how far this research will go in the future.

CHAPTER 2

Risk Factors Contributing to Alzheimer's Disease: A Comprehensive Investigation

2.1 Alzheimer’s Disease (AD) and its Impacts

Alzheimer's disease (AD) is a neurological condition that affects between 60 to 80% of all occurrences of dementia [5]. There are presently 50 million people suffering from dementia worldwide, however by the year 2050, 139 million people are anticipated to experience this condition due to rising life expectancy rates [6], [7] which will have a significant negative socioeconomic impact and affect the health care system [8] . The buildup of amyloid plaques and neurofibrillary tangles(NFTs) in the brain is a hallmark of Alzheimer's disease. The development of Alzheimer's disease, however, may also involve neuroinflammation, which is driven mostly by activated neuroglial cells, neutrophils, and macrophages [9]. In the beginning the entorhinal cortex and hippocampus, two areas of the brain crucial for memory, suffer damage. Later, it impacts the regions of the cerebral cortex in charge of social behavior, language, and cognition.. Damage ultimately spreads to other areas of the brain as well [10].

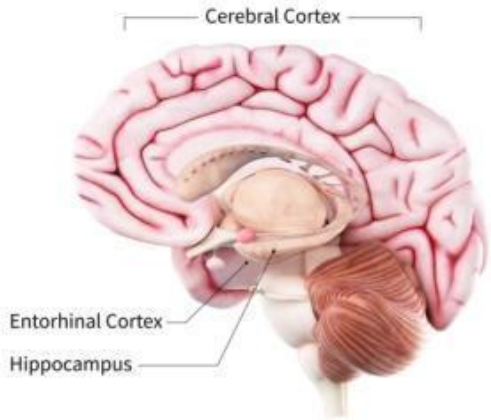

Figure 1: Structure of Brain

2.2 Symptoms

Loosing memory is the major symptom of Alzheimer's disease. Having problem with cognitive abilities is an alerting signal. But as the condition progresses, memory loss and other symptoms start to show themselves. When first diagnosed, a person with the

condition may already be experiencing the loss of memory and problems with mental clarity. As the symptoms grow, friend or family member may become more cognizant of the issues.

2.3 Memory Loss

Everyone occasionally experiences memory loss, but Alzheimer's disease is distinguished by continuous and deteriorating memory losss. The ability to perform daily duties at home or at work gradually suffers from memory loss.

- Repeating questions and remarks repeatedly is a symptom of Alzheimer's disease.
- Disregard all meetings, events, and interactions.
- Items are misplaced, frequently being deposited in odd places.
- Wander off in areas they used to know well.
- After a while, a person loses the names of their loved ones and commonplace items.
- Struggles to participate in conversations, convey ideas, or describe things using the appropriate words.

2.4 Behavior and Personality Changes

Alzheimer's disease-related brain alterations effect on emotions and behavior are possible. Any of the following might play a role in a problem:

- Depression.
- a decrease in the desire to engage in activities.
- Society's exclusion.
- Changes in emotion.
- Trust issues with others.
- Hostility.
- Changes in sleep habits, wandering.

2.5 Causes of Alzheimer’s Disease

Alzheimer's disease is characterized by persistent and acquired impaired memory over time, cognitive deficiencies in areas including language, spatiotemporal orientation, and executive function, along with behavioral shifts that all contribute to a gradual

degradation of personal autonomy [11]. In terms of histopathology, AD is distinguished by two pathognomonic features (Fig. 2) [12] : (1) the excessive intracellular accumulation of phosphorylated Tau protein that promotes the development of neurofibrillary tangles (NFTs) in subcortical gray matter and cerebral cortex; and (2) Neuritic plaques are extracellular Amyloid-beta peptide (A) fibril clumps. (NPs; Fig 2b) [13]. In the present scenario, it has been hypothesized that endogenous "damage signals," like Aβ oligomers, could activate microglial cells, releasing pro-inflammatory cytokines in the process. This would set off signaling cascades in neurons that would lead to tau protein hyperphosphorylation and aggregation. When neurons die, this protein is produced, which causes microglial cells to become activated and sets off a cyclic degenerative process that results in neurodegeneration [14] [15]. Because of this, NPs and NFTs are both involved in a number of neuronal processes that ultimately result in neuronal death [15] [16], synaptic alterations, oxidative stress, mitochondrial disturbances, neuroinflammation, changes in the permeability of the blood-brain barrier (BBB), and dysfunction of the neurovascular unit [16] [17].

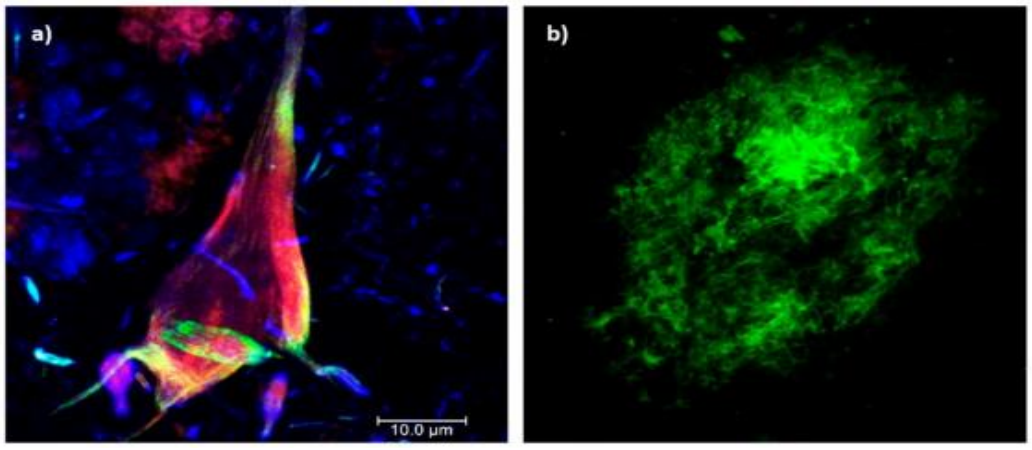

Figure 2: The histopathological signs of Alzheimer

Figure (a) Triple immunofluorescence image of a neurofibrillary tangle (conformational change: green channel; C-terminal tail: red channel; intact tau protein N-terminal: blue channel). A plaque made by amyloid is visible in [11] (b) immunofluorescence (Aβ 1–40: green channel). 100X photomicrographs with a calibration bar of 10 m [12].

CHAPTER 3

LITERATURE REVIEW

3.1Overview-Background Study

There are several approaches that are being explored for the Alzheimer's disease detection using AI. One method is to use machine learning algorithms to analyze brain imaging data, such as MRI or PET scans, to identify patterns or bio-markers that are associated with the disease. This can help to diagnose the disease earlier and more accurately than with traditional methods. Another approach is to use natural language processing techniques to analyze changes in a person's speech and language patterns, which can also be indicative of Alzheimer's disease.

Additionally, researchers are looking into the use of wearable technology, such as smart watches, to monitor changes in a person's daily activity and behavior that may be indicative of the disease. There is ongoing research into the use of AI for the detection of Alzheimer's disease. Other studies suggest that AI could be a beneficial aid for early detection of Alzheimer's disease. Popular deep learning techniques for text classification include Long-Short-Term Memory (LSTM) networks, Recurrent Neural Networks (RNNs), and Convolutional Neural Networks. Ensemble Learning: This approach uses multiple models and combines them to create a single model with greater predictive accuracy. Examples of ensemble learning techniques for text classification include bagging, boosting, and stacking. Rule- Based Learning: This approach uses manually-defined rules and heuristics to classify text. Rule-based systems can be hand-crafted by experts for specific domains, or generated automatically based on data. The aim of this paper is to improve the accuracy using BERT text classification method. The use of BERT text classification technology can significantly improve the accuracy and speed of early detection of Alzheimer's disease.

There are now two main areas of research into the diagnosis of AD in a person's spontaneous speech. One option is manual feature extraction,3 which can involve auditory features [18] [19], idiomatic expressions [20] [21] [22] [23], or a combination of the two [20]. More expert expertise is required for this procedure, and its success is not guaranteed. They are typically tied to a particular work scenario; when that scenario

shifts, the model's low universality becomes apparent because its artificially constructed characteristics and previous settings are no longer applicable. Deep learning is the alternative approach that can automatically pull out complex semantic information. In most cases, deep learning outperforms the traditional approach because of the strength of its representation learning capability. Furthermore, deep learning enhances the classifiers' generalization capabilities, allowing them to be used in a wider variety of clinical settings. Cascaded data from several non-linear processing units allows a deep neural network to automatically extract deep semantic characteristics through the process of representation learning.

3.2 AD detection using Deep Learning

Alzheimer's disease can be identified and diagnosed using deep learning, a branch of Artificial Intelligence. Numerous researchers have used deep learning techniques [23] [24] [25] [26] including RNN, LSTM networks (like ELMo [27]), and CNN, to identify AD in spoken speech. The hidden output of Bi-LSTM and word embedding are then made available for categorization using Recurrent Convolutional Neural Networks (RCNN) [28], which use them to collect contextual data. DPCNN [29] is a straightforward 15-layer network-model that mimics a deep CNN. It improves CNN's network depth without raising computing costs. Deep learning uses Medical image analysis using convolution neural networks (CNNs), such as brain MRI scans to recognize patterns in the brain which are indicative of Alzheimer’s disease. CNNs are also used to identify genetic markers specific to Alzheimer’s disease and to analyze other data sources. In 2018, researchers utilized a deep learning algorithm called a convolutional neural network (CNN) for the detection of Alzheimer's disease. The study analyzed over 1,000 brain MRI scans to evaluate how accurately the model could differentiate between healthy and degenerative brain scans [30].The results from the study showed that the CNN was able to accurately diagnose Alzheimer's with an accuracy rate of 92%. Deep learning algorithms can also be used to predict the progression of Alzheimer’s Disease.

Researchers have used various deep learning models, such as auto-encoders and recurrent neural networks, to predict the future trajectory of Alzheimer’s using data from brain scans and genetic markers. Overall, deep learning has been used successfully

to detect and diagnose Alzheimer's Disease with high accuracy. Deep learning models have the potential to provide earlier and more reliable diagnoses which can lead to improved medical care for patients. Additionally, deep learning can be used to identify novel risk factors for Alzheimer's Disease which may help prevent its occurrence. Researchers have utilized deep learning algorithms to identify brain structure-based biomarkers that are predictive of an individual's risk of developing Alzheimer's disease in the future.

Overall, deep learning is a rapidly developing area of AI research and holds promise for improving the detection, diagnosis, and prevention of Alzheimer's. However, further research is still needed to confirm the accuracy of deep learning models and to determine how best to use them in the clinical setting. Furthermore, deep learning can be used to identify new treatments for Alzheimer's Disease. Deep learning models can be trained to detect patterns in patient data and then used to recommend personalized treatments for each individual's unique condition. Overall, deep learning has proven to be incredibly useful for aiding in the detection and diagnosis of Alzheimer's as well as identifying potential treatments. Deep learning has the potential to revolutionize how Alzheimer's is managed in the future and could provide doctors with a powerful tool to help improve outcomes for their patients. In addition to its potential applications in diagnosis and treatment, deep learning can also be used to improve the quality of life of those affected by Alzheimer's.

Deep learning algorithms can also be used to detect changes in behavior or cognitive decline in individuals living with Alzheimer's, enabling healthcare workers to intervene early and provide support when necessary. Overall, deep learning holds tremendous potential for aiding in the diagnosis and management of Alzheimer's Disease in the future. The accurate and personalized insights generated by deep learning algorithms can be invaluable for providing better care for those affected by the disease.

As deep learning technology continues to evolve and improve, it is likely that the number of applications for deep learning in Alzheimer's research and treatment will continue to grow. Deep learning and deep transformer networks have been used to predict Alzheimer's disease. [31] research studied 165 papers that were written between 2005 and 2019 in total. Support vector machines (SVM), artificial neural networks (ANN), and deep learning (DL) were the three main machine learning methods that they concentrated on. SVM-based strategies were more reliable, in their analysis, and they hoped that deep learning techniques might produce better outcomes in the future.

The system was composed of three components: an encoder layer, a GPNet (Gaussian Process-based Neural Network) layer, and a SoftMax classifier. The authors used an independent dataset to evaluate the performance of their system. Results demonstrated that the model achieved significantly higher accuracy than traditional machine-learning methods and a commercially available baseline systems. The work shows the potential of utilizing deep learning architectures for Alzheimer's Disease (AD) detection. Other research shows that deep learning-based approaches can also be used to diagnose and detect AD in natural language processing (NLP) applications.

3.3AD detection using Transformers

Automated methods for diagnosing Alzheimer's disease (AD) have been made possible by recent developments in natural language processing (NLP). More recently, transformer-based techniques have been proposed for the diagnosis of AD from text. Transformer-based methods are a type of deep learning architecture that have been used in NLP tasks such as language modeling, machine translation, and text classification (Vaswani et al., 2017) [32]. The Transformer encoder was successfully used with GP-Net to detect Alzheimer's disease (AD) across a variety of datasets. The ADReSS, Pitt, and iFLy datasets were specifically used for the evaluation. The system's durability and generalizability were impressively highlighted by its excellent accuracy on each dataset. The algorithm correctly detected AD cases in the ADReSS dataset with an accuracy of 74.3%. Its greater performance was demonstrated by its even higher accuracy of 91.4% on the Pitt dataset. The model also demonstrated 81.6% accuracy on the iFLy dataset, further demonstrating its capacity to detect AD successfully across several datasets. These encouraging findings highlight the Transformer encoder with GP-Net's potential as a useful tool for AD detection research [4]. Scientists outside the realm of medicine have given this investigation into the changes in language functions brought on by AD a lot of attention [33]. Researchers, particularly those in the field of natural language processing, have suggested computer-based methods for automatic and semiautomatic language analysis in AD patients [34] [35] [36] [37] [38] [39] [40] [41].

An ensemble of T1-weighted MRI scans sliced along the coronal axis is employed in this study to diagnose AD, and the Bottleneck Transformers deep learning model

(Srinivas et al., 2021) is used as the base classifier with a sharpness aware minimizer [42]. The categorization method used by one study in 2015 and another in 2020 in their separate research publications has been replicated here [43] [44]respectively. Future work can be done on improving the performance of transformer-based methods for the detection of AD from other types of data such as images and videos.

Two different auto-regressive LSTM-based neural network language models were used by researcher Fritsch et al.[23] to classify AD and HC transcripts using the Pitt corpus from the DementiaBank dataset. Then, to work on predicting AD, Pan et al. [45] utilized stacked bidirectional LSTMs and gated recurrent unit (GRU) layers equipped with a hierarchical attention mechanism. The overall model includes the GloVe word embedding sequence.

3.4 Data Set

The data read step in deep learning is the process of loading data into the model from a data-set. This step is important for training the model as it requires the availability of a large amount of data in order to generate accurate predictions. The data set contains text from both AD and non-AD persons from Public Pitt available on Kaggle. The text is based on their observations after viewing a picture. This data can be used to analyze the differences in how AD and non-AD persons perceive the same image. It can also be used to identify patterns in how different people interpret visual information. This data set can also be used to compare the cognitive abilities between those with Alzheimer's Disease and those without, as well as to investigate how different types of visual information are processed by different people. This data-set consists of text of different persons based on their observation after showing them a picture. This data is of AD, non-AD persons [46].

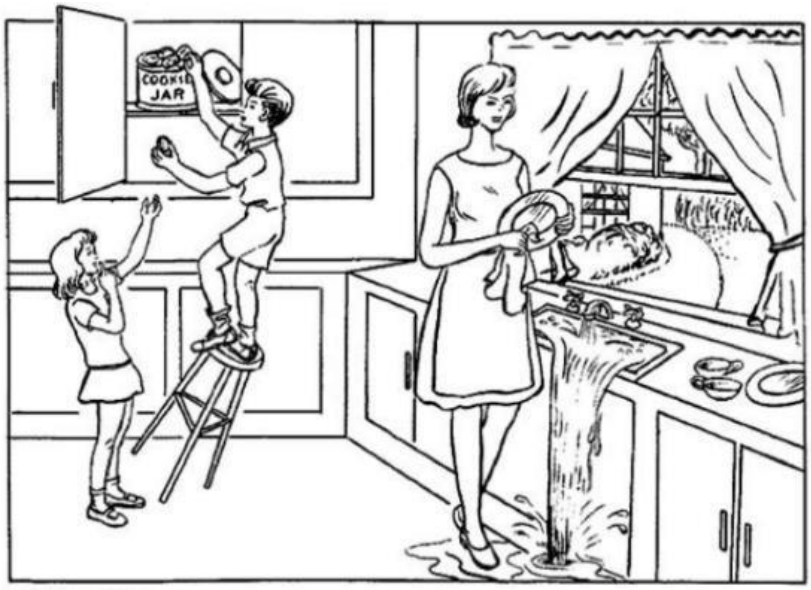

Figure 3: Reference Image

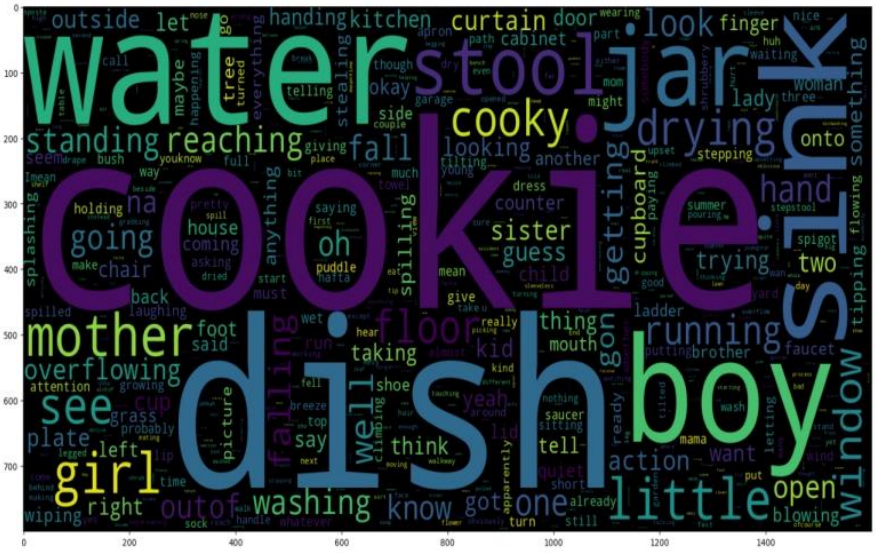

Figure 4: Dataset analysis

The dataset contains a variety of features and columns that offer important data for our investigation. However, we included data augmentation approaches to increase the dataset's diversity and strengthen our model's generalization abilities.

CHAPTER 4

4.1 Proposed Methodology

4.1.1 BERT

Bidirectional Encoder Representations from Transformers, or BERT, is a ground-breaking method in the study of natural language processing (NLP). It has been effectively used for a variety of NLP tasks, including but not limited to sentiment analysis, named entity recognition, question answering, and text categorization. BERT has also shown useful in specific fields, such the detection of Alzheimer's illness.
BERT is fundamentally a "transformer" model. The deep learning model architecture known as transformers completely changed the NLP industry. Transformers, in contrast to older NLP models, can analyze and comprehend text by taking into account the relationships between words in a certain context. BERT makes use of this transformational architecture to improve word and phrase comprehension in a more sophisticated and nuanced way.

4.1.2 What makes BERT efficient?

In order to pretrain bidirectional representations, BERT [23] builds on Transformer networks [33] by incorporating both left- and right-side conditioning concurrently in all levels. By determining whether a sentence in the input corresponds to a certain sentence in the corpus or not, and by predicting randomly masked words in the input, the representations are simultaneously optimized. According to the BERT authors, bidirectionality enables the model to quickly adjust for a downstream purpose with little architecture modification. The method for a number of NLP benchmarks was significantly improved by BERT [45] [47].

BERT's bidirectional nature is one of its important characteristics. BERT takes into account both directions concurrently, in contrast to earlier models that processed text in a left-to-right or right-to-left manner. This implies that BERT considers how each word fits in relation to the terms that come before and after it when analyzing a list of

words in a phrase. BERT can better understand the meaning of the text by taking into account the entire context. This allows it to pick up on more nuanced word dependencies and nuances.

BERT has substantially enhanced the study of NLP by utilizing the strength of transformers and bidirectional analysis. It has made significant advancements in a variety of language-related activities, including the recognition of disorders like Alzheimer's, thanks to its capacity to understand the intricate meanings of words and phrases. BERT has remained a key model in the NLP field, advancing the field's comprehension of natural language and assisting in the creation of smarter, context-aware systems.

- ## Self-Attention:

The BERT (Bidirectional Encoder Representations from Transformers) model relies heavily on self-attention. It is essential for identifying the contextual relationships between words in a given text sequence. Self-attention enables BERT to analyze all word dependencies and relationships simultaneously rather than sequentially or in a predetermined order.

Using the Transformer design, which comprises of numerous layers of feed-forward and self-attention neural networks, the self-attention mechanism is implemented in BERT. Let's examine how BERT handles self-attention:

Word embeddings that correspond to the words in the input text are used as the BERT's input. Each word's meaning and contextual information is encoded into a high-dimensional vector representation.

Self-Attention: The Transformer's layers all engage in self-attention. The words in the sequence are broken down into three vectors called the Query (Q), Key (K), and Value (V) vectors. The learnt weight matrices are multiplied by the word embedding to produce these vectors.

Attention Scores: Attention scores are calculated for each word in the sequence by taking the dot product between the word's Query vector and the Key vectors of all the other words in the sequence. The importance or relevance of each word in relation to the current word is indicated by these attention scores.

To acquire **attention weights**, the attention scores are scaled and then put via a softmax function. The contribution of each word to the representation of the current word during the attention process is determined by these weights. Stronger relationships are implied by higher attention weights.

Contextual Representation: Weighted representations of each word are created by applying the attention weights to the Value vectors. The contextual representation of the current word, encompassing information from every other word in the sequence, is created by adding together these weighted representations.

Multi-Head Attention: BERT makes use of a number of attention heads, each of which is a separate self-attention mechanism. The different dependencies and patterns in the text are captured by each attention head, giving a more complete picture of the input sequence.

Self-attention and feed-forward neural networks are **stacked in multiple layers** in the BERT to record progressively more complicated contextual data. The contextual representations obtained from the preceding layer are improved by each subsequent layer, allowing the model to capture long-range dependencies and complex interactions between words.

By making use of self-attention, BERT is able to gather both local and global contextual data, which enables it to produce detailed and context-aware word representations. These representations are essential for a number of later natural language processing tasks, including question answering, named entity recognition, and text classification.

4.1.3 Pre-process

Pre-processing is an important part of many NLP tasks, such as for BERT. Pre-processing is a set of steps that prepare data for further analysis and downstream tasks. In the context of BERT, pre-processing steps include tokenization, adding special tokens, masking words, and segmenting pieces of text. Tokenization is the first step in the pre-processing pipeline. It involves identifying each word or piece of text in the input sentence and separating them into individual pieces. A tokenizer is used to do this, so that each token is associated with an index based on its position in the sentence. Addition of unique tokens like [CLS] and [SEP] is the following step.

Every sentence has the [CLS] token added to the beginning to denote the beginning of the sentence, and the [SEP] token is added to denote the end of the sentence. This helps the model to classify sentences during the training process.

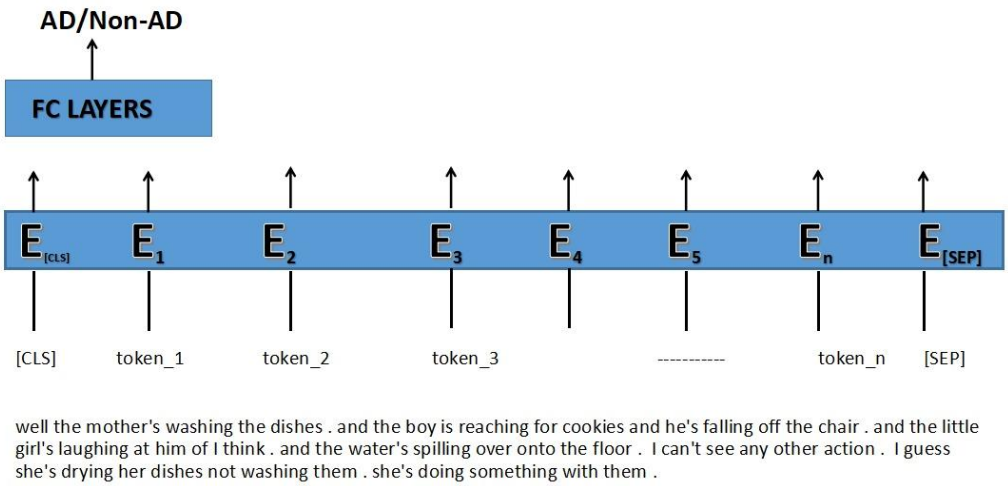

Figure 5: Preprocess

Following this, the next step is to mask words in the sentence. Masking is the process of randomly replacing certain words in the sentence with a [MASK] token.

The purpose of this is to ensure that the model does not overfit to the training data, since randomly masking words will force the model to learn the context of words in the sentence, instead of memorizing them. Finally, the last step of the pre-processing pipeline is to segment pieces of the text. This involves separating a sentence into two or more smaller pieces, which can then be fed into the model as separate inputs. By segmenting pieces of text, the model is able to learn longer dependencies, since it can wrap its attention around long pieces of text. Overall, pre-processing is an essential part of training the BERT model, as it prepares the data so that it can be used in the training process.

Pre-proces steps such as tokenization, adding special tokens, masking words, and segmenting pieces of text are all necessary for the model to learn accurately, as these steps ensure that it does not over fit to the training data. In addition to the steps mentioned above, pre-processing can also involve other tasks such as stemming and lemmatization. Stemming is the process of reducing words to their root form (e.g. "talked" is reduced to "talk"), while lemmatization is the process of transforming words into their base form (e.g. "better" is transformed to "good"). These processes can help

reduce the number of tokens that are inputted into the model and make them easier to learn. Pre-processing is a key component of many natural language processing tasks, and for BERT it is no different. Without the pre-processing steps mentioned above, the model would not be able to accurately learn the context of words and sentences, as these steps help ensure that it does not overfit to the training data. Therefore, pre-processing is essential for the success of BERT.

Pre-processing can take some time, but it is a necessary step in the BERT model training process. To ensure that the model is accurately trained, you should make sure to apply all the key pre-processing steps, such as tokenization, adding special tokens, masking words, segmenting pieces of text, and applying stemming and lemmatization. Doing so yields better results for your BERT model, since all these steps prepare the data for accurate learning by the model.

4.1.4 Encoder

Once the pre-processing steps are complete, the next step is to feed the data into the BERT model. This is done through the use of an encoder, which is a component of the BERT model that is responsible for transforming the input data into a meaningful representation. The encoder takes in the tokenized input data and uses a multi-layer The model architecture's key component, the encoder component of BERT is in charge of parsing incoming text and producing contextualized representations. BERT's encoder, which is based on the transformer architecture, has several layers of feed-forward and self-attentional neural networks. It captures contextual dependencies through a bidirectional analysis of input tokens. Each token can attend to itself while also attending to others, assessing their relative importance.

Feed-forward networks improve representations further. Multiple layers are stacked to allow BERT to capture intricate relationships in the text. Each token's contextualized representations are produced by the encoder and are improved by the context in which they are used. The subsequent NLP tasks can benefit from these representations. Overall, the encoder in BERT uses feed-forward networks and self-attention to create contextualized token representations, which allows the model to comprehend & interpret natural language text successfully.

4.2 Optimizer

An optimization technique called an optimizer algorithm can help a deep learning model perform better. These optimization methods or optimizers have a substantial impact on the deep learning model's efficiency and training speed. But initially, the query about what an optimizer is appears.

As you train the deep learning optimizer model, adjust the weights for each epoch and lower the loss function. A process or technique known as an optimizer alters the weights and learning rates of a neural network. As a result, it helps to improve precision while reducing overall loss. Since deep learning models generally have millions of parameters, choosing the right weights for the model can be difficult.

It emphasizes how crucial it is to choose an optimization algorithm that is suitable for your application. Therefore, before diving into these machine learning techniques, data scientists must understand them.

Gradient Descent (GD): Gradient Descent is the most basic but most used optimization algorithm. Algorithms for categorization and linear regression both widely utilize itThe backpropagation of neural networks also employs gradient descent. Gradient descent is a first-order optimization technique that relies on the first order derivative of a loss function. In order for the function to reach a minima, it dictates how the weights should be modified. Backpropagation is used to pass the loss from one layer to the next, and the weights of the model are changed to reflect the losses in order to lessen the loss.

Stochastic Gradient Descent (SGD): The SGD optimization approach adjusts the parameters in the direction of the loss function's negative gradient. A small batch of training sample samples are processed at a time to compute the gradients.

Adam (Adaptive Moment Estimation): Adam short form of Adaptive Moment Estimation is a stochastic gradient descent extension that modifies the rate at which each parameter learns based on estimations of the first and second moments of the

gradients. Due to its success in a range of deep learning tasks, it is widely employed.

Adagrad: Adagrad (Adaptive Gradient), a method of optimization, modifies the learning rate for each parameter based on prior gradients. For infrequent values, it accelerates learning rate while slowing it down for frequent parameters.

Adadelta: Adagrad's aggressive and monotonically declining learning rate is addressed by Adadelta, an extension of Adagrad. The learning rate is dynamically adjusted depending on a shifting window of gradient updates.

RMSprop: Another optimization approach that adjusts the learning rate based on the moving average of the squared gradients is RMSprop (Root Mean Square Propagation). Deep learning model convergence is accelerated as a result.

For our models we used Adam optimizer as it best fits the model. Due to its adjustable learning rates and momentum, Adam optimizer is renowned for its efficiency in optimizing deep neural networks. Because of its adaptability, it may dynamically alter the learning rate for each parameter, promoting faster convergence and better handling of sparse gradients. By effectively adjusting the model's weights and biases during the training phase with the aid of the Adam optimizer, we can enhance performance overall. Adam is the greatest option for obtaining the finest outcomes because of its established track record and compliance with our model architecture.

4.3 Training Testing Split

The available data-set is typically split into two independent sets: a training set and a testing set, in order to reliably and impartially assess the model's performance. With the help of the training set, the model is developed, and the testing set is used to evaluate its performance.

This division serves to guarantee that the model is tested using data that it was not exposed to during training. We may evaluate the model's propensity to generalize and make predictions on fresh, untested data by utilizing a separate testing set.

In this particular case, the data-set was divided in such a way that 20% of the data was allocated for testing, while the remaining 80% was used for training. The training set, which comprises 80% of the data, is utilized to train the model on the task at hand. During training, the model learns patterns and relationships in the data, enabling it to make predictions or classifications.

Additionally, both the training and testing sets must have a representative distribution of data from various classes or categories. By ensuring that the model learns from and is evaluated on a wide variety of samples, the dataset's many patterns and classifications may be understood and predicted more thoroughly.

4.4 Data Augmentation

With the help of various text-based modifications, a method known as data augmentation for text data can be utilized to fictitiously increase the training dataset. These modifications bring about changes in the text, increasing its diversity and enhancing the model's capacity to generalize to various inputs. Word replacement, random insertion, random deletion, random swapping, and random character deletion/insertion are examples of common procedures. Word replacement includes changing specific words with synonyms or comparable words, whereas random insertion involves adding new words to the text. Random deletion eliminates words to represent missing or omitted content, and random swap shuffles neighboring words to highlight the significance of word order. In order to imitate faults or noise found in real-world writing, random character deletion/insertion alters individual characters inside words.

These methods can be used to improve the dataset and make NLP models more robust, able to handle text variations and perform better overall.

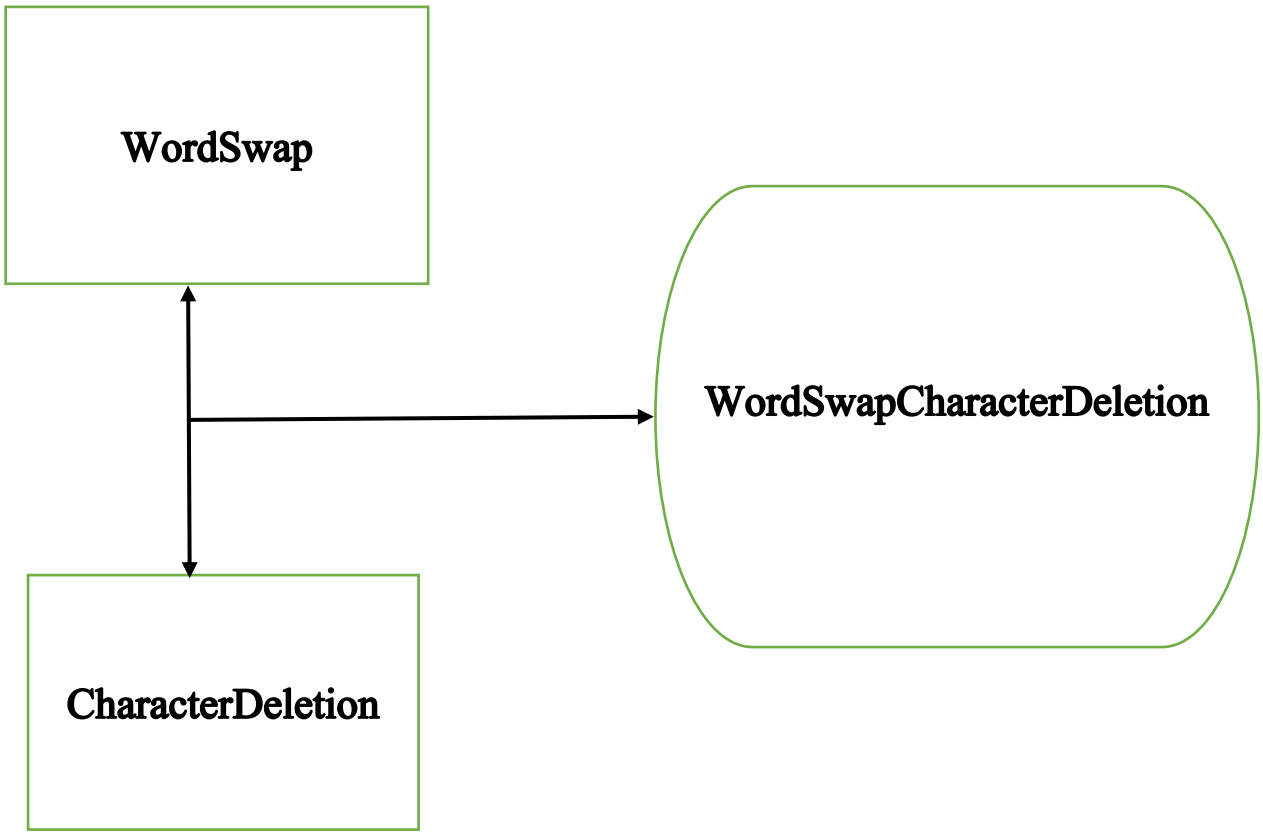

Figure 6: Data Augmentation technique

4.5 Neural Network

The layers of the models are built using the algorithm of machine learning for various problems of regression and classification. Multiple layers of interconnected nodes, commonly known as neurons, make up neural networks. With regard to the network's learning process, each layer has a distinct function. Our model uses the following layers and activation functions.

Sequential mode: Sequential Model: A sequential model is a stack of layers that are added one after the other in a linear fashion. It is the most straightforward and typical technique to construct neural networks, particularly when working with feedforward networks or straightforward sequential structures.
When utilizing the Sequential model, you may add layers to the network by defining the layer type and its parameters using the.add() method. From the input layer to the output layer, the data moves sequentially via the layers. For networks with a single input and one output, this method works well.
For applications like picture classification or sequence classification, where the data flows through the network without significant branching or merging, sequential models are frequently used.

Input Layer: The initial data or features for the neural network are received by the input layer. One neuron is provided for each input feature, representing the problem's input dimensions.

Hidden Layers: Between the input and output layers, there exist hidden layers. They work with the incoming data through computations, gradually extracting higher-level features. According to the problem's difficulty, each layer's number of neurons as well as the total number of hidden layers might vary.

Activation Functions: By introducing non-linearity to the neural network, activation functions let it learn and approximately understand complicated relationships in the data. The **"Sigmoid"** function, "ReLU (Rectified Linear Unit)", **"tanh (hyperbolic tangent)"**, and **"Softmax"** (used in the output layer for multi-class classification) are examples of common activation functions.

- **Sigmoid:**The sigmoid function—also referred to as the logistic function—is a nonlinear activation function that converts input values into a range between 0 and 1. Effective modeling of non-linear connections is made possible by its distinctive S-shaped curve. The sigmoid function can reduce the output to a value that resembles probability, making it particularly useful for binary classification applications. Its vanishing gradient issue for big input values, however, can

obstruct the training process.

- **ReLU:** Popular activation function known as the Rectified Linear Unit (ReLU) solves the vanishing gradient issue. If the input is positive, it outputs the value straight; if not, it outputs zero. Deep learning models frequently use the ReLU function since it is computationally effective. ReLU makes neural networks learn more quickly by adding sparsity to the activations, especially in conditions where sparse features are common. The "dying ReLU" phenomenon, where neurons may become irreversibly dormant during training, is a potential drawback of ReLU.
- **Tanh:** Another popular activation function that transfers input values to a range between -1 and 1 is the hyperbolic tangent (tanh) function. Tanh displays an S-shaped curve, much like the sigmoid function. In hidden layers of neural networks, the tanh function is helpful since it normalizes the inputs. It is appropriate for applications where the data exhibits symmetry about zero since it may record both positive and negative values. For large input values, it can, like the sigmoid function, experience the vanishing gradient problem.

- **Softmax:** It creates a probability distribution across classes from a vector of real numbers. The outputs are normalized by the softmax function, which guarantees that the probabilities add up to one. It makes decision-making easier and enables the selection of the most likely class by presenting class probabilities. The softmax function can, however, create unbalanced probabilities when working with unbalanced datasets since it is susceptible to outliers. The activation function softmax is used in the final layer of the neural network for the purpose of classification of multi-class.

4.6Recurrent Neural network:

Recurrent neural networks (RNNs) were first proposed as a cognitive model of incremental language processing (Elman, 1990). RNNs model inputs that take the form of sequences. Since RNN language models' architecture includes a "memory" component that stores data from earlier time-steps, they theoretically have the ability to model sequences of arbitrary length. This desirable property allows RNN

language models to be unrestricted by the markov assumption. The hidden state at a specific timestep, denoted as ht, is computed by incorporating the current input at that timestep, represented as xt, which typically corresponds to a dense representation of the word or feature at that time. Additionally, the previous timestep's hidden state, ht-1, is used as part of the computation to generate the current hidden state. A feed-forward network with dynamic hidden layers is how an RNN might be conceptualized. Since the decision made by the model at a particular time-step is jointly influenced by "memory" from prior time-steps provided by the hidden layer and the input from the current time-step, RNNs can theoretically represent extended sequences of free text.

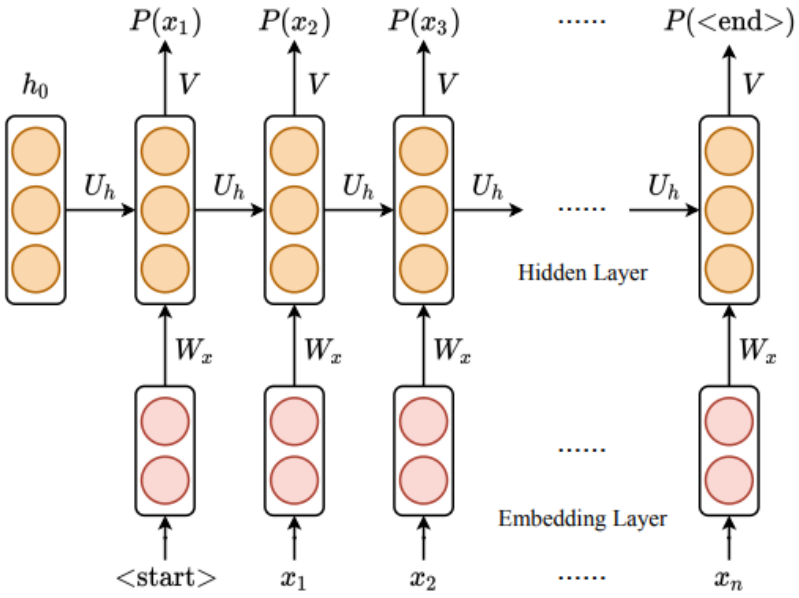

Figure 7: RNN

The problem of disappearing or exploding gradients arises frequently while optimizing RNNs over lengthy sequences. The weight matrix, Uh, is continuously multiplied by the hidden state h throughout training. The gradients of each hidden state at each time-step with respect to the loss are multiplied by the same number repeatedly, proportional to the length of the sequence, during backpropagation. As a result, the gradients either grow very large (explode), or they are pushed towards zero (vanish), which makes it difficult to learn or represent the sequence.

There are further methods designed expressly to deal with these problems. Here are several methods frequently applied in relation to RNNs:

Gated Recurrent Units (GRU) and Long Short-Term Memory (LSTM)

A particular RNN architecture known as a GRU uses gating methods to regulate the information flow within the network. The update gate (z) and the reset gate (r) are its two principal gates.

Update Gate (z): Manages how much of the current state should be combined with the new candidate activation and how much of the previous hidden state should be kept.

Reset gate: Determines how much of the previous concealed state should be ignored in order to take into account the fresh input through the reset gate (r).

The network may update and spread information over time in a selective manner because to the architecture of the GRU. By retaining important information and preventing irrelevant information from gathering and amplifying over time, it lessens the vanishing gradient issue.

Another RNN variant, **LSTM**, tackles the issue of vanishing gradients and captures long-term relationships by utilizing a more intricate gating mechanism. The input gate (i), forget gate (f), and output gate (o) are the three gates that are introduced.

- *Entry Gate (i)* controls how much fresh data should be stored in the memory cell.
- *The Forget Gate (f)* establishes how much previously recorded information should be maintained or forgotten.
- *Output gate (o)* controls the quantity of data that will be output from the memory cell is the output gate.

A memory cell that can preserve data across lengthy sequences is introduced by LSTM. The network can handle gradients and capture long-term dependencies more effectively thanks to the memory cell's ability to read, write, and erase information selectively.

Mathematically LSTM is formulated as:

$$f_t = \sigma(x_t W_f + h_{t-1} U_f + b_f$$

$$i_t = \sigma(x_t W_i + h_{t-1} U_i + b_i$$
$$o_t = \sigma(x_t W_o + h_{t-1} U_o + b_o$$
$$\hat{C}_t = tanh(x_t W_c + h_{t-1} U_c + b_c)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \hat{C}_t$$
$$h_t = o_t \odot \tanh C_t$$

Gradient clipping: In RNNs, gradient clipping can be used to reduce the magnitude of gradients during backpropagation, similar to how it is used in other neural networks. By downscaling gradients that go over a certain limit, this avoids the ballooning gradient problem.

Truncated Backpropagation Through Time (BPTT): Truncated BPTT restricts the number of time steps across which gradients are backpropagated, as opposed to propagating gradients through the full sequence. This makes the optimization process more manageable by lowering the possibility of gradients vanishing or exploding.

Initialization Methods: The vanishing gradient issue can be resolved by properly initializing recurrent weights. More stable gradients can be produced via initialization techniques like orthogonal initialization or recurrent-specific initialization techniques like the Identity matrix initialization.

4.7RCNN

Recurrent convolutional neural networks (RCNNs) are a particular form of neural network architecture used to analyze sequential input, such as text or time series data, by combining convolutional layers and recurrent layers. The input sequence's local and temporal dependencies are also targets of this algorithm.

- **Convolutional Layers:** These layers use convolutional operations on the input sequence to extract features. To discover regional patterns and characteristics and to capture spatial information inside the sequence, convolutional filters are used.
- **Recurrent Layers:** By keeping a hidden state that is updated as the

network analyzes each piece in the sequence, these layers are in charge of modeling temporal dependencies. The network can keep track of data from earlier time steps and incorporate it into the present forecast thanks to the recurrent connections.

- **Pooling Layers:** These layers are frequently used to decrease the dimensionality of the output features after the convolutional and recurrent layers. With less computational complexity, pooling makes it possible to combine the derived characteristics and capture higher-level representations.

Table 1: Concordance between predicted label and ground truth

	True Class	
Predicted Title	Positive	Negative
Positive	True Positive(TP)	False Positive(FP)
Negative	False Negative(FN)	True Negative(TN)

Embeddings:

To represent words, sentences, or documents as numbers, numerous embedding types are employed in natural language processing (NLP) and machine learning. Here are some prevalent embedding types:

- With **one-hot encoding**, every word is represented as a sparse binary vector, with a value of 1 where the word's index should be and a value of 0 everywhere else.
- Distributed Representations: **Word2Vec**: Based on the context in which words appear, it uses shallow neural networks to produce dense word embeddings.
  GloVe (Global Vectors for Word Representation) factors word co-occurrence

information to create word embeddings.

- **FastText:** Extends Word2Vec by handling non-vocabulary words by representing words as bags of character n-grams.
- A variant of Word2Vec called **Paragraph Vector** (Doc2Vec) learns fixed-length embeddings for variable-length texts, such sentences or documents.
- A pre-trained model called **the Universal Sentence Encoder (USE)** converts sentences into fixed-dimensional vectors that include both syntactic and semantic information.
- **Transformers' BERT** (Bidirectional Encoder Representations): a transformer-based approach that takes the complete input context into account to produce contextualized word and phrase embeddings.
- **ELMO (Embeddings from Language Models)** uses a deep bidirectional language model to generate contextualized word embeddings.
- **GPT (Generative Pre-trained Transformer):** A transformer-based approach that creates contextualized embeddings at the word or sentence level using unsupervised learning.

4.8Implementation

A CSV dataset file served as the basis for our investigations in this study. The dataset's numerous features and columns give us important data for our investigation. However, we included data augmentation approaches to increase the dataset's diversity and strengthen the generalization abilities of our model. We want to introduce variability and improve the resilience of our dataset by undertaking this augmentation. As a feature extractor in our methodology, a pre-trained BERT stands for Bidirectional Encoder Representations from Transformers model was used. Modern language models like BERT extract detailed contextual information from text data. The BERT encoder is used to create contextualized embeddings, which represent the semantic meaning of each word or subword token, from the input text.

We preprocess the data after collecting the BERT embeddings to make it appropriate for further investigation. To maintain uniform input dimensions across all samples, this preprocessing stage entails operations like tokenization, padding, and truncation.

The preprocessed BERT embeddings are then fed into a neural network architecture to produce the output. Depending on the requirements of the particular task, the neural network is made by different layers, which include convolutional layers, recurrent layers, and fully connected layers. These layers take in high-level representations of the incoming data by learning and extracting pertinent features from the BERT embeddings. By combining the power of BERT contextual embeddings with the adaptability of a neural network, our strategy allows the network to learn complex patterns and relationships within the data, increasing performance on the target task.

4.9Ensemble

Average ensembling is a powerful technique used in machine learning to improve the performance of predictive models. To build a more precise and reliable ensemble model, it includes merging the predictions of various independent models. Average ensemble can successfully reduce bias and variation by taking advantage of the variety of predictions made by many models, improving overall performance. Because it lessens the effects of outliers or specific model biases, the average ensemble is robust. Due to their unique architectures or training methods, several models may excel at collecting various parts of the data or display various biases. The ensemble model can successfully smooth out these biases and produce more balanced forecasts by averaging their results. This robustness helps to lessen over-fitting and enhance generalization, which is especially useful when working with noisy or ambiguous data.

It is crucial to choose a diverse selection of models that have been trained on the same dataset or problem in order to implement an average ensemble. To ensure variety in their predictions, the models should differ in terms of their architectures, algorithms, hyperparameters, or training procedures. Once each model has been trained and its predictions have been produced, it is possible to average the forecasts by adding the expected values and dividing them by the total number of models. Before averaging, additional transformations or adjustments to the forecasts may be required depending on the particular issue.

4.10 Algorithm

1. Initialize the model with input data. (augmented data file)
2. Load the pre-trained BERT model.
3. Use BERT Preprocess for tokenization, padding sequence of text input.
4. Use BERT Encoder to extract semantic features from text input.
5. Create the layers of the model.
6. Training the model.
7. Compute the loss between the predicted labels and ground truth labels using a binary-cross entropy loss function.
8. Predicting model accuracy on test data.
9. Update the model parameters using an optimizer.

- **Flow Chart**

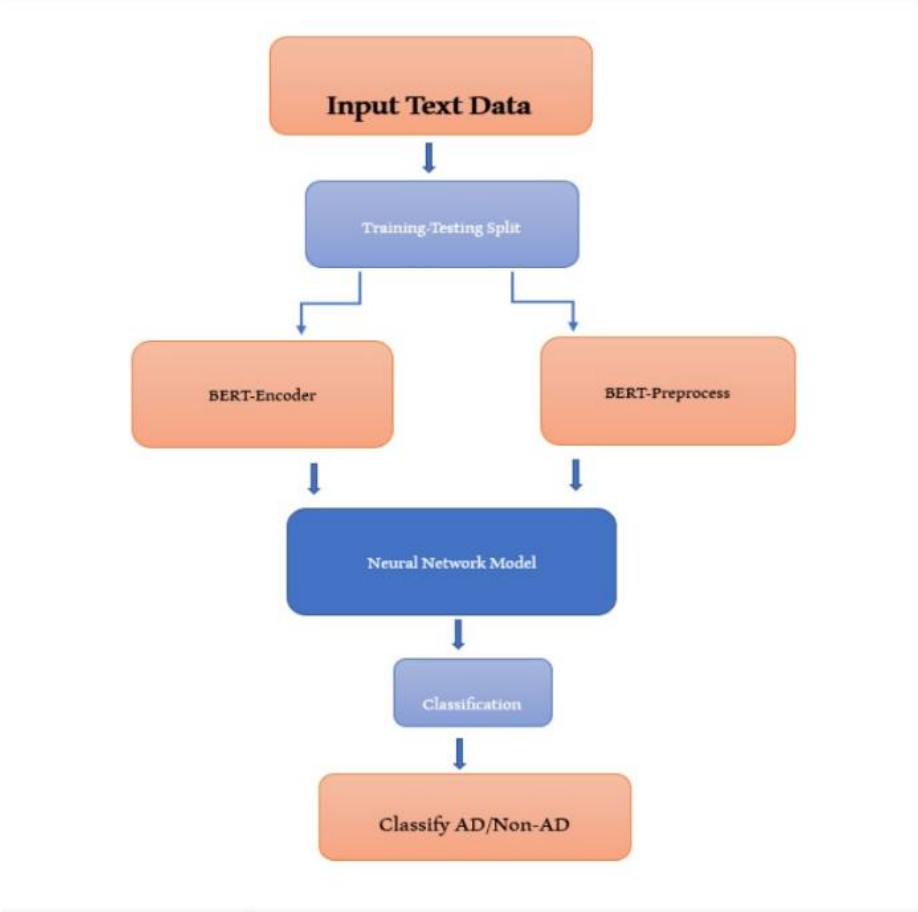

Figure 8: Flow chart

4.11 Block diagram:

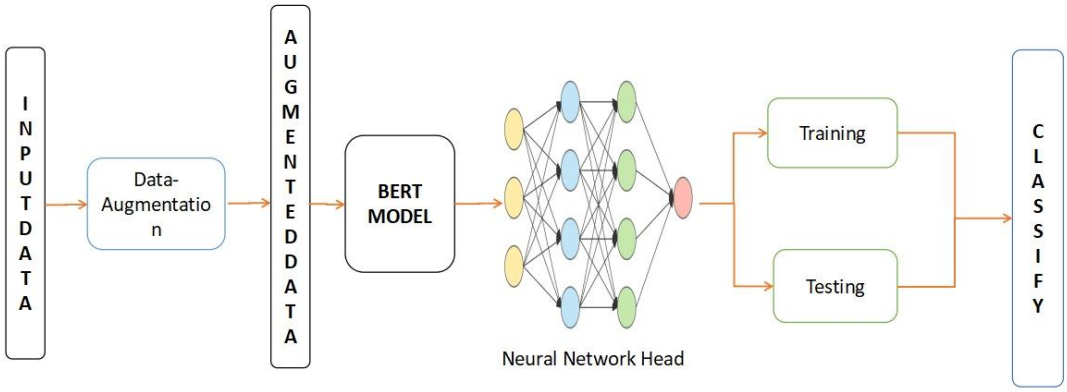

Figure 9: Block diagram

4.12 Experiment Results:

The data-set was split into ratios of 20,80 for testing and training respectively. 20% of the portion was utilized for testing, while 80% was used for training. The model shows the performance measures for the model once the results from the completed epochs cycle & model is compiled. Accuracy, precision, recall, and F1 score were evaluation measures for classification that were based on the correspondence among the real and expected classes.

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FP} + \text{FN} + \text{TP}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{NP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Table 2: Results on PublicPitt DementiaBanks

Model	**Classifier**	**Precision**	**Accuracy**	**Recall**
BERT	**CNN**	**89.47**	**83.16**	**72.34**
BERT	**RCNN**	**89.94**	**93.33**	**83.20**
BERT(Augmented-data)	**CNN**	**79.54**	**74.16**	**71.72**
BERT(Augmented-data)	**RCNN**	**97.08**	**95.99**	**95.49**
Ensemble	**BERTCNN + BERT-RCNN**	**-**	**94.98**	**-**

The best that we can tell, compared to earlier models on the Pitt datasets, the present model's success shows considerable performance improvements. This accomplishment shows the possibilities of merging BERT models with other methods and demonstrates the effectiveness of synergistic methods for natural language processing problems. The present model performs better than prior ones, indicating that it has successfully addressed the drawbacks of those earlier methods and proving its supremacy in handling the challenging Pitt datasets. This accomplishment has important ramifications because it shows improvements in NLP research and possible uses in areas including sentiment analysis, text classification, and information retrieval. The present model's higher performance also demonstrates the possibility for ongoing innovation and advancement in the creation of NLP models, pushing the limits of what is possible in tasks requiring natural language interpretation. The results highlight the value of benchmarking and assessing models using standardized datasets, like the Pitt datasets, in order to provide a fair and unbiased assessment of their capabilities.

4.13 Analyzing Graph Data for Technical Insights:

Following are the outputs graphs of the implemented models:

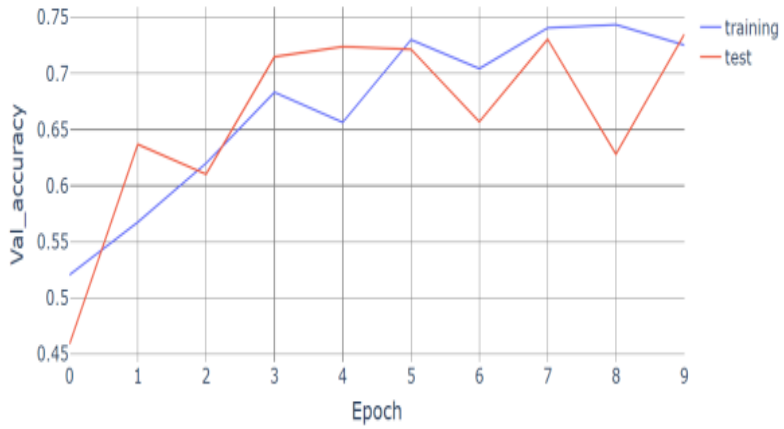

Figure 10: CNN Accuracy

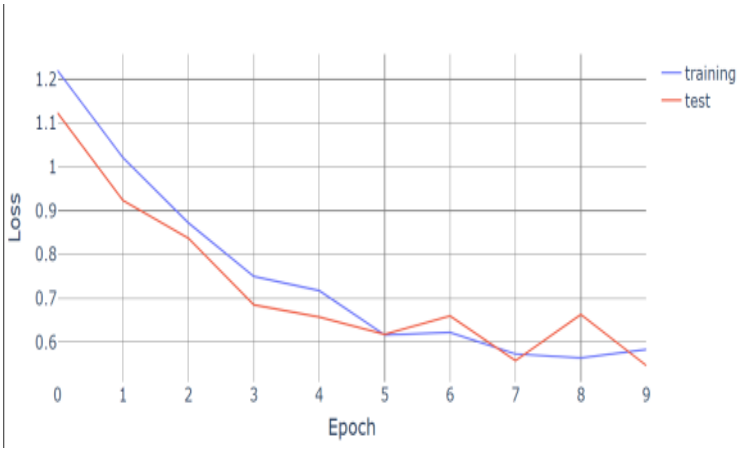

Figure 11: CNN Loss

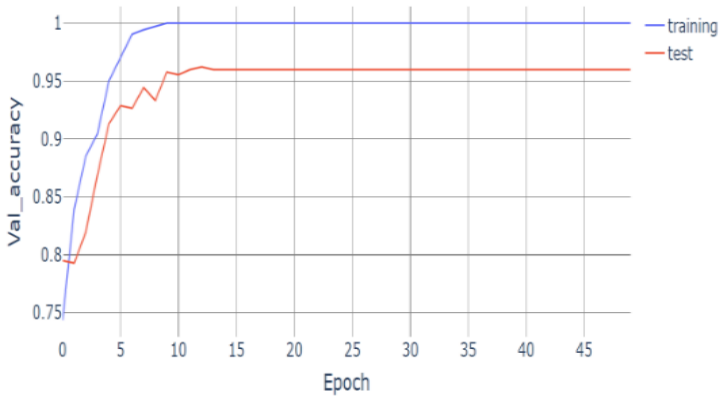

Figure 12: RCNN Accuracy

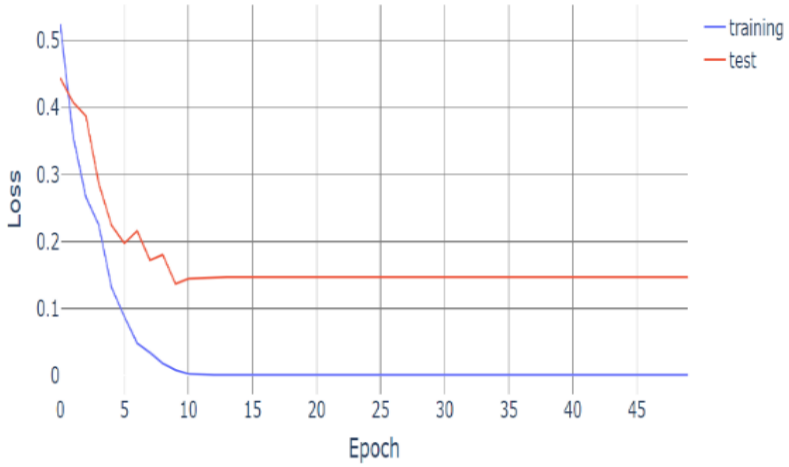

Figure 13: RCNN Loss

The above graphs display the accuracy and loss metrics for two models: BERT-CNN and BERT-RCNN, both trained on augmented data input files. The results indicate that BERT-CNN achieved an accuracy of 0.7416 and loss 0.5, while BERT-RCNN achieved a higher accuracy of 0.9599 and loss 0.1465.

These metrics are commonly used to evaluate the performance of machine learning models. Accuracy measures the proportion of correctly classified instances out of the total number of instances. In this case, BERT-CNN correctly classified 74.16% of the augmented data, while BERT-RCNN achieved a higher accuracy of 95.99%.

Loss, on the other hand, represents the discrepancy between the predicted outputs and the actual outputs. It is typically used as an optimization objective for training the model. Lower loss values indicate a better fit to the data. Unfortunately, the loss values for both models are not provided in the information given.

It's important to note that the choice between BERT-CNN and BERT-RCNN depends on the specific task and dataset. While BERT-RCNN seems to have achieved higher accuracy in this case, other factors such as model complexity, training time, and interpretability should also be considered when selecting the appropriate model for a given task. Also, other parameters such as precision and recall are also important performance parameters in evaluating the model. We can acquire a more thorough insight of the model's performance in text classification by taking precision and recall into account in addition to accuracy.

CHAPTER 5

5.1 Discussion

The findings of this investigation are really encouraging. The usage of BERT with CNN and BERT with RNN models greatly improved accuracy when compared to the preceding methods. Combining the contextualized embeddings power of BERT with CNN and RNN architectures allowed us to better perform on our classification task and extract more nuanced information from the input text. Convolutional layers were used in the BERT with CNN model to extract regional features and patterns from the BERT-encoded text representations. The model was able to gather the most pertinent data across the entire sequence thanks to the following pooling layer. Dense layers were added to the model to further enhance the retrieved features and help it develop discriminative representations for the classification challenge. The BERT-RCNN architectural combination showed success in gathering both local and global contextual data, leading to increased accuracy.

A recurrent layer, especially LSTM, was used in the BERT with RNN model to simulate the sequential dependencies found in the BERT-encoded text representations. Long-range dependencies and contextual information were easier to simulate across the whole input sequence thanks to the LSTM layer. The encoded representations were then combined by the pooling layer, and the learned features were then further improved by dense layers. By combining BERT with an RNN architecture, the model was able to capture temporal dynamics and contextual information with more efficiency and accuracy. Results shows that Bert is a powerful tool for detecting Alzheimer disease and could be used in clinical settings to help diagnose and treat patients. Additionally, this research could be used to develop more accurate algorithms for detecting other diseases as well. Further research should be conducted to explore the potential applications of Bert in medical diagnosis and treatment. It is essential to be mindful when interpreting the findings of this study. The accuracy of the algorithm may vary depending on the data-set used and the specific parameters used in the model. Additionally, further research should be conducted to evaluate the performance of Bert on other datasets and in different contexts. Overall, this study provides evidence that Bert is a powerful tool for detecting Alzheimer disease and could be used in clinical settings to help diagnose and treat patients. Further research should be conducted to explore the potential applications of Bert in medical diagnosis and treatment.

Table 3: Comparison of Classification Scores on Pitt datasets

Method	**Embedding**	**Classifier**	**Precision**	**Recall**	**Accuracy**
Sweta Karlekar [21]	POS	CNN-RNN	-	-	91.1
Fritsch et al. [23].	n-grams	NNLM+LSTM	-	-	85.6
Orimaye et al. [29]	n-grams	D2NN	-	-	88.9
Fraser et al. [46]	35 Hand-Crafted Feature	LR	-	-	81.92
Yancheva et al. [48]	12 Cluster-Based Features + LS&A	Random Forest	80.00	80.00	80.00
Sirts et al [49]	Cluster+PID+SID Features	LR	74.4 ±1.5	72.5 ±1.2	-
Hernandez et al. [50]	105 Hand-Crafted Features	SVM	81.00	81.00	79.00
Roshanzamir et al. [51]	BERT Base (Sentence Level)	LR	90.31 ±7.36	76.52 ±8.06	84.46 ±6.31
Roshanzamir et al. [51]	Bert Large	LR	90.57 ±3.1	84.34 ±7.58	88.08 ±4.48
Pan et al. [45]	GloVe Word Embedding Sequence	BiLSTM|GRU Hierarchical Attention	84.02	84.97	-
Li et al. [52]	185Hand-Craft Features	LR	-	-	77
Fraser et al. [53]	Info and LM Features	SVM	-	-	75
Transformer FP[25] [4]	Transformer +Feature projection	Transformer	88	91	75
Transformer + GP [4]	Transformer + Feature purification	Transformer	94	89	93.5
Model 1	**BERT**	**CNN**	**89.47**	**72.34**	**83.16**
Model 2	**BERT**	**RCNN**	**89.94**	**93.33**	**83.20**
Model 3	**BERT**(Augmented-data)	**CNN**	**79.54**	**71.72**	**74.16**
Model 4	**BERT**(Augmented-data)	**RCNN**	**97.08**	**95.49**	**95.99**
Model 5	**Ensemble**		-	-	**94.98**

Bold text represents the results of models implemented in this paper, where last three rows show results on an augmented dataset.

CHAPTER 6

6.1Conclusion and Future Work

The model consists of a pre-trained BERT encoder and a classifier [20]. The BERT encoder is used to encode the text data into a fixed-length vector, which is then fed into the classifier to predict whether a patient has AD or not. The model is trained on a corpus of clinical notes from patients with AD and healthy control subjects. The results show that the proposed model performs well other state-of-the-art models for the early diagnosis of AD from text. The use of BERT text classification to detect Alzheimer's disease has outperformed other deep learning models using tokenizer activation function layers and optimizer function. This is due to the fact that the BERT model is able to capture the long- range dependencies between words in a sentence and its contextual information, which other models are not able to do. Furthermore, the BERT model is able to achieve this with fewer layers and fewer parameters, making it more efficient. Finally, BERT's use of the transformer architecture makes it more robust to changes in data and able to generalize better than other models. All these advantages make BERT an excellent choice for detecting Alzheimer's disease. A growing body of research has utilized artificial intelligence (AI) to accurately detect Alzheimer's disease (AD). AI-based methods have been proposed in recent years that enable the automated diagnosis of AD through a variety of approaches, including computer vision and natural language processing (NLP). By extracting clinically relevant features from neuroimaging data (MRI, PET) and analyzing text from medical notes and other sources, AI-based methods can generate reliable diagnostic models for AD. Specifically, AI-based computer vision analyses have achieved high accuracy in predicting AD pathology through the use of convolutional neural networks (CNNs) and deep learning algorithms trained on large image databases. Deep learning algorithms can also be used to detect changes in behavior or cognitive decline in individuals living with Alzheimer's, enabling healthcare workers to intervene early and provide support when necessary. Overall, deep learning holds tremendous potential for aiding in detection and management of AD in future.

The accurate and personalized insights generated by deep learning algorithms can be invaluable for providing better care for those affected by the disease. As deep learning technology continues to evolve and improve, it is likely that the number of applications for deep learning in Alzheimer's research and treatment will continue to grow.

Additionally, AI-based NLP algorithms such as BERT (Bidirectional Encoder Representations from Transformers) have demonstrated promising results in predicting AD from medical notes and free-text documents. These approaches could eventually be used in clinical practice to aid in the earlier detection of AD and other forms of dementia. Additionally, researchers are exploring the use of wearable technology, such as smartwatches, to monitor changes in a person's daily activity and behavior that may be indicative of the disease.

In conclusion, the encoder part of the BERT model is a deep learning model that uses Transformer architecture to encode a sequence of words into contextualized representations. The representations can capture long range dependencies and relationships which is useful for various NLP tasks. The decoder allows the model to be used for different tasks requiring various outputs.

References

[1] R. M. Sousa, C. P. Ferri, D. Acosta, E. Albanese, M. Guerra , . Y. Huang, . K. S. Jacob and A. T. Jotheeswaran, "Contribution of chronic diseases to disability in elderly people in countries with low and middle incomes: a 10/66 Dementia Research Group population-based survey," November 2009.

[2] H. Goodglass, E. Kaplan and B. Barresi, "Boston Diagnostic Aphasia Examination," in *Encyclopedia of Clinical Neuropsychology*, Austin, Pro-Ed, 2001, pp. 428-430.

[3] V. Taler and N. A. Philips, "Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review," pp. 501-556, June 2008.

[4] N. Liu, Z. Yuan and Q. Tang, "Improving Alzheimer's Disease Detection for Speech Based on Feature Purification Network," 2022.

[5] "2022 Alzheimer's disease facts and figures. (2022).," *Alzheimer's & dementia : the journal of the Alzheimer's Association*, , pp. 700-789, 2022.

[6] "GBD 2016 Dementia Collaborators (2019). Global, regional, and national burden of Alzheimer's disease and other dementias, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016," *The Lancet. Neurology*, 2016.

[7] "World Health Organization (WHO)," [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia. [Accessed 10 April 2023].

[8] K. A. Matthews, W. Xu, A. H. Gaglioti, J. B. Holt, J. B. Croft, D. Mack and L. C. McGuire, "Racial and ethnic estimates of Alzheimer's disease and related dementias in the United States (2015-2060) in adults aged ≥65 years," *the journal of the Alzheimer's Association*, 2019.

[9] L. A. C, J. Hardy and J. M. Schott, "Alzheimer's disease," *European journal of neurology*, 2019.

[10] "National Insitute on Aging (NIA)," [Online]. Available: https://www.nia.nih.gov/health/what-happens-brain-alzheimers-disease. [Accessed April 2023].

[11] J. Andrade-Guerrero , A. Santiago-Balmaseda, P. Jeronimo-Aguilar, I. Vargas-Rodríguez , A. . R. Cadena-Suárez and C. Sánchez-Garibay , "Alzheimer's Disease: An Updated Overview of Its Genetics," *International journal of molecular sciences*, 2023.

[12] R. B. Maccioni, J. P. Tapia and L. Guzman-Martinez , "Pathway to Tau Modifications and the Origins of Alzheimer's Disease," *Archives of medical research*, vol. 49, no. 2, pp. 130-131, 2018.

[13] A. González, S. K. Singh, M. Churruca and R. B. Maccioni, "Alzheimer's Disease and Tau Self-Assembly: In the Search of the Missing Link," *International journal of molecular sciences*, vol. 23, no. 8, 2022.

[14] M. V. F. Silva, C. d. M. G. Loures, L. . C. V. Alves, L. . C. d. Souza, . K. . B. G. Borges and M. d. G. Carvalho, "Alzheimer’s disease: risk factors and potentially protective measures," *Journal of Biomedical Science*, vol. 26, no. 33, 2019.

[15] A. B. Reiss, H. A. Arain, M. M. Stecker, N. M. Siegart and L. J. Kasselman, "Amyloid toxicity in Alzheimer's disease," *Reviews in the neurosciences*, vol. 29, no. 6, pp. 613-627, 2018.

[16] L. O. Soto-Rojas, M. Pacheco-Herrero, P. A. Martínez-Gómez, B. B. Campa-Córdoba , R. Apátiga-Pérez, M. M. V, C. R. Harrington, . F. d. l. Cruz, L. Garcés-Ramírez and . J. Luna-Muñoz, "The Neurovascular Unit Dysfunction in Alzheimer's Disease," *International journal of molecular sciences*, vol. 22, no. 4, 2022.

[17] D. Beltrami, G. Gagliardi, R. R. Favretti, E. Ghidoni, F. Tamburini and L. Calzà, "Speech Analysis by Natural Language Processing Techniques: A Possible Tool for Very Early Detection of Cognitive Decline?," *Frontiers in aging neuroscience*, vol. 10, 2018.

[18] K. C. Fraser, J. A. Meltzer and F. Rudzicz, "Linguistic Features Identify Alzheimer's Disease in Narrative Speech," *Journal of Alzheimer's disease*, vol. 49, no. 2, pp. 407-422, 2016.

[19] G. Gosztolya, V. Vincze, . L. Tóth, M. Pákáski, J. Kálmán and I. Hoffmann, "Identifying Mild Cognitive Impairment and mild Alzheimer’s disease based on spontaneous speech using ASR and linguistic features," *computer speech and language* , vol. 53, pp. 181-197, 2019.

[20] K. C. Fraser, K. . L. Fors and D. Kokkinakis, "Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment," *Computer Speech & Language,* vol. 53, 2018.

[21] S. Karlekar, T. Niu and M. Bansal, Detecting Linguistic Characteristics of Alzheimer's Dementia by Interpreting Neural Models, 2018.

[22] K. Lopez-de-Ipiña, J. B. Alonso, J. Solé-Casals, N. Barroso, P. Henríquez Rodríguez, M. Faundez-Zanuy, C. M. Travieso, M. Ecay, P. Martinez-Lage, U. Martinez-de-Lizarduy, H. E. Martinez and A. Ezeiza, "On Automatic Diagnosis of Alzheimer's Disease Based on Spontaneous Speech Analysis and Emotional Temperature," *Cognitive Computation,* vol. 7, 2013.

[23] J. Fritsch, S. Wankerl and E. Noeth, Automatic Diagnosis of Alzheimer's Disease Using Neural Network Language Models, 2019, pp. 5841-5845.

[24] K. Lopez-de-Ipina, U. Martinez-de-Lizarduy, P. M. Calvo, . B. Beita, J. García-Melero, M. Ecay-Torres, A. Estanga and M. Faundez-Zanuy, "Analysis of Disfluencies for automatic detection of Mild Cognitive Impartment: a deep learning approach," *2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI),* pp. 1-4, 2017.

[25] F. D. Palo and N. Parde, "Enriching Neural Models with Targeted Features for Dementia Detection," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 2019.

[26] M. E. Peters, M. Neumann, M. Iyyer and M. Gardner, "Deep contextualized word representations".

[27] S. Lai, L. Xu, K. Liu and J. Zhao, "Recurrent Convolutional Neural Networks for Text Classification," *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence 2267,* pp. 2267-2273, 2015.

[28] R. Johnson and T. Zhang, Deep Pyramid Convolutional Neural Networks for Text Categorizatio, 2017.

[29] S. O. Orimaye, J. S.-M. Wong and C. P. Wong, "Deep language space neural network for classifying mild cognitive impairment and Alzheimer-type dementia," *PLOS ONE,* vol. 13, 2018.

[30] S.-H. Wang, Y. Zhang, Y.-J. Li, W. Jia, F.-Y. Liu, M.-M. Yang and Y. Zhang, "Single slice based detection for Alzheimer's disease via wavelet entropy and multilayer perceptron trained by biogeography-based optimization," 2018.

[31] N. Singh, D. Patteshwari, N. Soni and A. Kapoor, "Automated detection of Alzheimer disease using MRI images and deep neural networks- A review," 2022.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, "Attention is All you Need," *Advances in Neural Information Processing Systems 30 (NIPS 2017),* 2017.

[33] F. Rudzicz, L. C. Currie, A. Danks, T. Mehta and S. Zhao, "Automatically identifying trouble-indicating speech behaviors in Alzheimer's disease," *ASSETS14 - Proceedings of the 16th International ACM SIGACCESS Conference on Computers and Accessibility,* 2014.

[34] A. Satt, R. Hoory, A. König, P. Aalten and P. Robert, Speech-Based Automatic and Robust Detection of Very Early Dementia, 2014.

[35] K. Lopez-de-Ipiña, U. Martinez-de-Lizarduy, N. Barroso, M. Ecay, P. Martinez-Lage, F. Torres and M. Faundez-Zanuy, Automatic analysis of Categorical Verbal Fluency for Mild Cognitive impartment detection: A non-linear language independent approach, 2015.

[36] A. Khodabakhsh and C. Demiroglu, "Analysis of Speech-Based Measures for Detecting and Monitoring Alzheimer's Disease," *Methods in molecular biology (Clifton, N.J.).*

[37] L. Padró and E. Stanilovsky, FreeLing 3.0: Towards Wider Multilinguality, Istanbul, Turkey: European Language Resources Association (ELRA), 2012.

[38] J. Lyons, "python_speech_features 0.6," [Online]. Available: https://pypi.org/project/python_speech_features/. [Accessed 15 March 2018].

[39] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing,* vol. 14, 2004.

[40] L. Breiman, "Random Forests," *Machine Learning,* 2001.

[41] M. Asgari, J. Kaye and H. H. Dodge, "Predicting mild cognitive impairment from spontaneous spoken utterances," *Alzheimer's & Dementia: Translational Research & Clinical Interventions 3(2),* 2017.

[42] A. Srinivas, T.-Y. Lin, N. Parmar and J. Shlens, "Bottleneck Transformers for Visual Recognition,"

2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[43] C. Salvatore, A. Cerasa, P. Battista and M. C. Gilardi, "Magnetic Resonance Imaging biomarkers for the early diagnosis of Alzheimer's Disease: a machine learning approach," *Frontiers in Neuroscience* .

[44] D. Pan, A. Zeng, L. Jia, Y. Huang, T. Frizzell and X. Song, "Early Detection of Alzheimer's Disease Using Magnetic Resonance Imaging: A Novel Approach Combining Convolutional Neural Networks and Ensemble Learning.," *Frontiers in Neuroscience*, 2020.

[45] Y. Pan, B. Mirheidari, M. Reuber, A. Venneri, D. J. Blackburn and H. Christensen, "Automatic Hierarchical Attention Neural Network for Detecting AD," *Interspeech*, 2019.

[46] J. T. Becker, F. Boller , O. L. Lopez , J. Saxton and K. L. McGonigle, "The natural history of Alzheimer's disease. Description of study cohort and accuracy of diagnosis.".

[47] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019.

[48] M. Yancheva and F. Rudzicz, Vector-space topic models for detecting Alzheimer's disease, 2016.

[49] K. Sirts, O. Piguet and M. Johnson, "Idea density for predicting Alzheimer's disease from transcribed speech," 2017.

[50] L. E. Hernandez-Dominguez, S. Ratté, G. Sierra and A. G. R. Bergua, "Computer-based evaluation of AD and MCI patients during a picture description task," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 2018.

[51] A. Roshanzamir, H. Aghajan and M. Soleymani, Transformer-based deep neural network language models for Alzheimer's disease detection from targeted speech, 2020.

[52] B. Li, Y.-T. Hsu and F. Rudzicz, Detecting dementia in Mandarin Chinese using transfer learning from a parallel corpus.

[53] K. C. Fraser, N. Linz, B. Li, K. L. Fors, F. Rudzicz, A. König, J. Alexandersson, P. H. Robert and D. Kokkinakis, Multilingual Prediction of Alzheimer's Disease Through Domain Adaptation and Concept-Based Language Modelling, 2019.