Comparative analysis of Medical Text Classification using Transformer

By

Rohit Chakraborty

(B.Tech 3nd Year, Enrollment No.:12021002001017)

Under the Supervision of

Prof SAGARIKA GHOSH

Dept. of Computer Science & Engineering

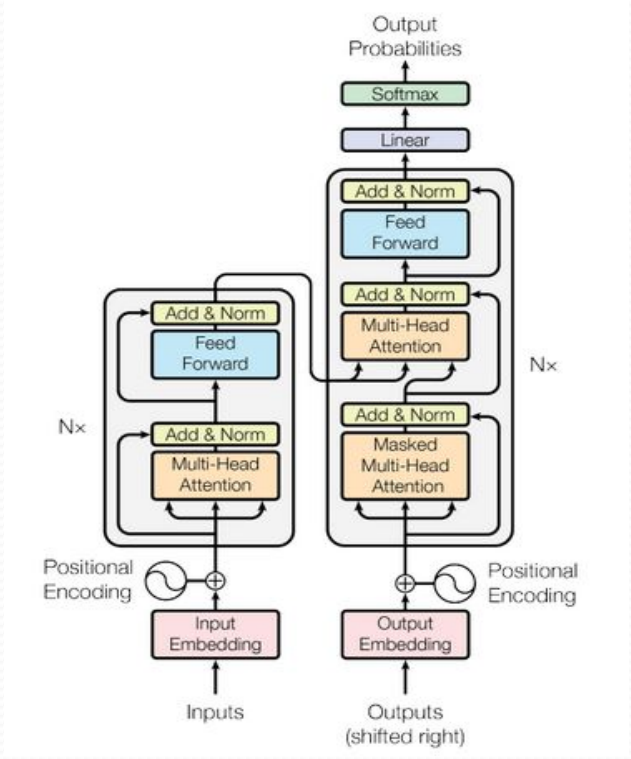
University of Engineering & Management, Jaipur

Outlines

- Introduction
- Literature Review
- Objectives
- Model Analysis
- Experimental Setup
- Result Analysis
- Limitations
- Conclusion & Future Scope Reference
- Acknowledgement

Introduction

Utilizing advanced deep learning methods in NLP for medical text classification was the primary focus of this research. The aim was to classify diseases from medical data using state-of-the-art models like LSTM, BERT, Roberta, GPT2, and a BERT+LSTM hybrid. The dataset encompassed five disease categories: tumors, gastrointestinal diseases, nervous system diseases, cardiovascular diseases, and general pathological conditions. Our methodology involved employing transformer-based pre-trained models and fine-tuning strategies to optimize their performance in disease classification. These transformer models, known for their ability to comprehend intricate linguistic nuances, were incorporated into the classification process.



Literature Review

1. Medical-Based Text Classification Using FastText Features and CNN-LSTM Model Mohamed Walid Zeghdaou
2. Outpatient Text Classification Using Attention-Based Bidirectional LSTM for Robot-Assisted Servicing in Hospital
3. News Text Classification Based on Improved Bi-LSTM-CNN
4. FasTag: Automatic text classification of unstructured medical narrative
5. Turkish Medical Text Classification Using BERT
6. A GPT-2 Language Model for Biomedical Texts in Portuguese
7. BioGPT: generative pre-trained transformer for biomedical text generation and mining

Objectives

The main objective is to employ advanced deep learning planning methodologies to precisely categorize a medical text dataset, specifically aimed at identifying diseases within medical abstracts. This process entails the utilization of Transformers sourced from the Hugging Face Library as the core model, followed by fine-tuning using supplementary layers to enhance classification effectiveness. These Transformers encompass models such as BERT, RoBERTa, GPT-2, LSTM, and a hybrid model like BERT+LSTM.

The utilization of a transformer head enables the model to comprehend the linguistic nuances within the text and generate embeddings. These embeddings serve as valuable inputs for the fine-tuning layers, aiding in the classification process.

Proposed Model

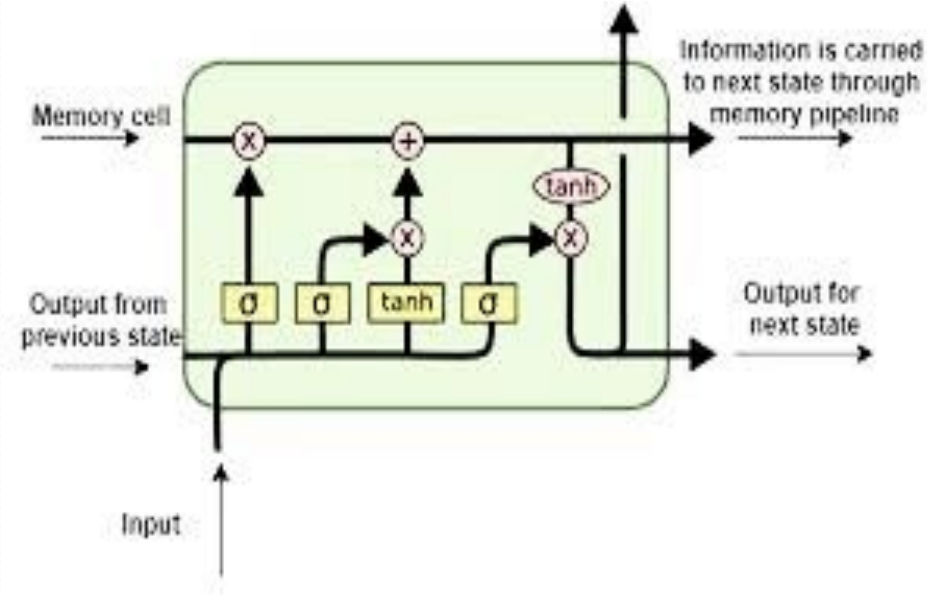
BERT+LSTM model:

- The bert-base-uncased model is used To attend the bidirectional linguistic features of the text Using the attention mechanism and create embeddings
- Then the embeddings are transferred by a dropout layer which then reduces the dimensionality to prevent from overfitting
- Then it is send. Then it is sent to a LSTM layer with 256 units.This layer the dependencies and features in a sequential manner for text sequence classification

Experimental Set-up

LSTM:

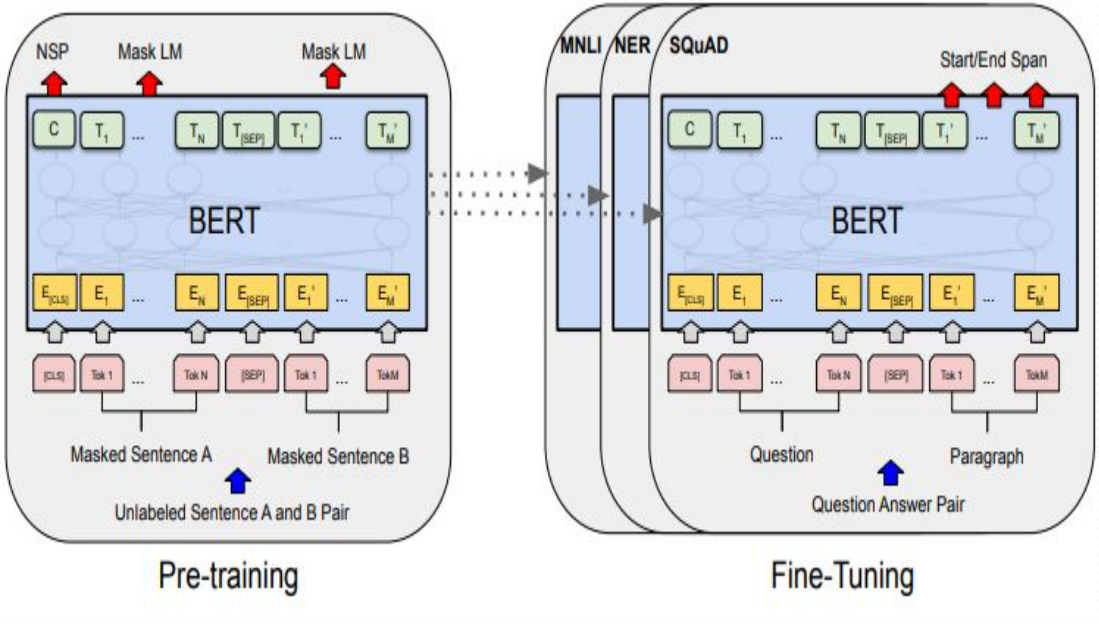
- The model constitutes a 512 memory unit for understanding complex features of the medical abstract
- With a dropout layer of 0.5 units and classifier layer with 5 units with softmax activation function



Experimental Set-up

BERT-BASE-UNCASED:

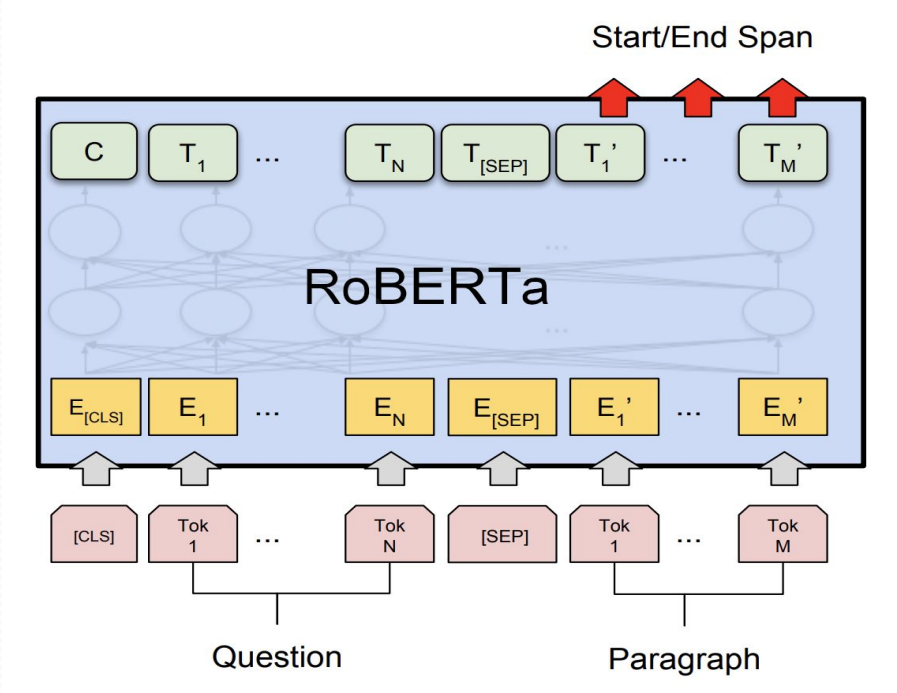
- The base model of bert which is used to understand the linguistic features of the text data and create embeddings
- Then passed on to and dropped layer, and then a classifier layer to classify the data



Experimental Set-up

RoBERTa:

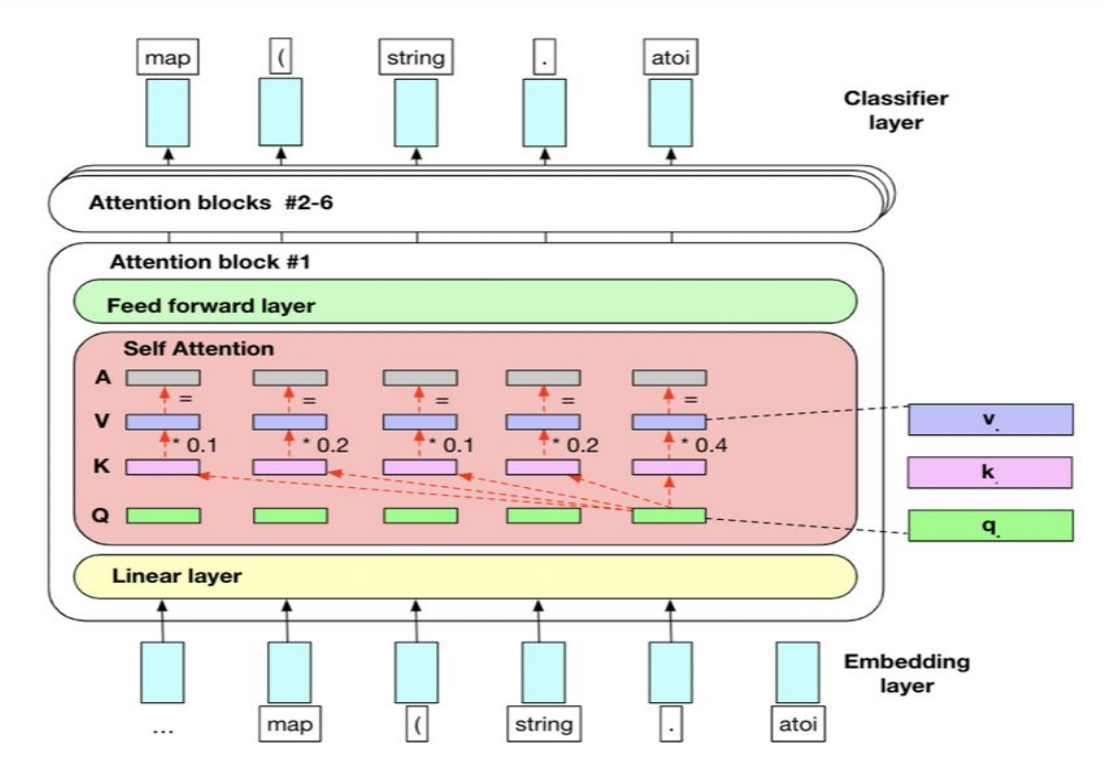
- The base model of roberta which is used to understand the linguistic features of the text data and create embeddings
- Then passed on to and dropped layer, and then a classifier layer to classify the data



Experimental Set-up

GPT-2:

- The base model of GPT-2 utilized for its contextual comprehension of textual data
- Then passed on to and dropped layer,Additionally, a classifier is integrated into the architecture, enabling the model to categorize and classify medical text based on the learned representations



Result Analysis

MODEL	TRAIN ACCURACY(%)	TEST ACCURACY(%)
LSTM	72	52
BERT	90	70
ROBERTA	87	72
GPT-2	67	62
BERT+LSTM	91	71

Limitation

1. The transform models are very resource hungry models due to which they are difficult to train for long epochs as they take lot of time in training
2. They also need a lot data for finetune purpose which was unavailable due which there has been problem of overfitting during training

Conclusions & Future Scope

1. From the training of the models, we can infer that using the BERT+LSTM model can give better accuracy with more training .
2. The future shape of the model is to use it for name recognition And a multimodal data set with more accuracy and data
3. And make an app for it

References

1. "Medical Abstracts Dataset | Papers With Code".
2. Outpatient Text Classification Using Attention-Based Bidirectional LSTM for Robot-Assisted Servicing in Hospital
3. Medical-Based Text Classification Using FastText Features and CNN-LSTM Model

Acknowledgement

I would want to convey my heartfelt gratitude to Prof. Sagarika Ghosh, my mentor, for her invaluable advice and assistance in completing my project. She was there to assist me every step of the way, and her motivation is what enabled me to accomplish my task effectively. I would also like to thank all of the other supporting personnel and HOD CSE department Prof Mrinal Kanti Sarkar for who assisted me by supplying the equipment that was essential, without which I would not have been able to perform efficiently on this project. I would also want to thank the University of Engineering and Management for accepting my project in my desired field of expertise. I'd also like to thank my friends and parents for their support and encouragement as I worked on this assignment

Dept. of CSE, University of Engineering & Management Jaipur