Midterm Project

Performance of explainer in different text classification models

Zexin Ren

March 22, 2022

Outline

Introduction

Explainer

"Why Should I Trust You?": Explaining the Predictions of Any Classifier

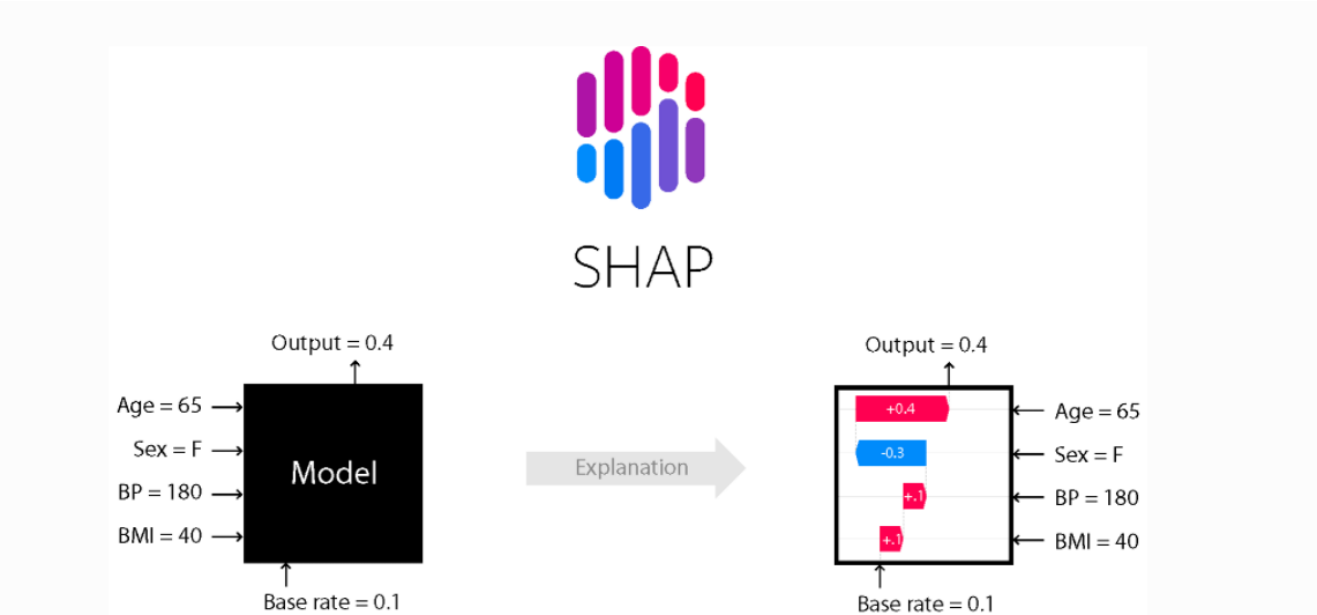


Figure: An example of SHAP explainer

Introduction

Model and Dataset

	Model 1	Model 2
Num. of Labels	2	5
Model Name	distilbert-base-uncased	distilbert-base-uncased
Tokenizer Name	distilbert-base-uncased	distilbert-base-uncased
Dataset	Clinical Statement	Medical abstracts
Test Accuracy	85.5%	77%

Method

Salience Andlysis

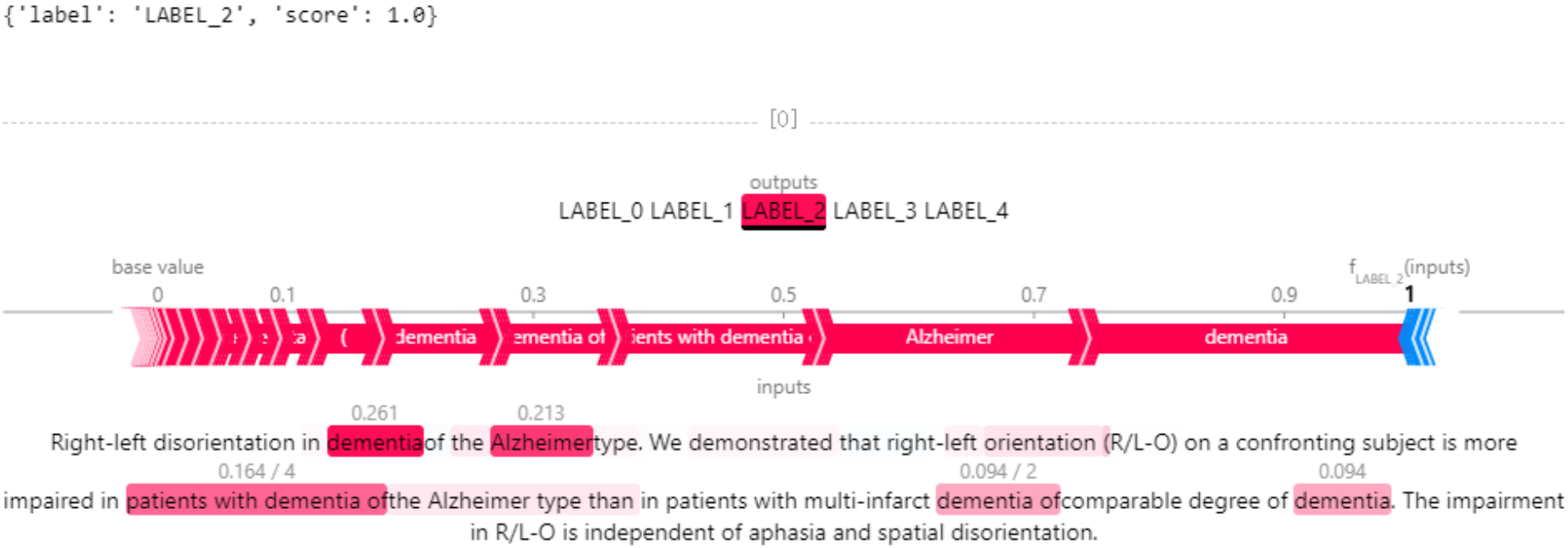


Figure: A Text Example

Method

Top K Mask

How to test this result?

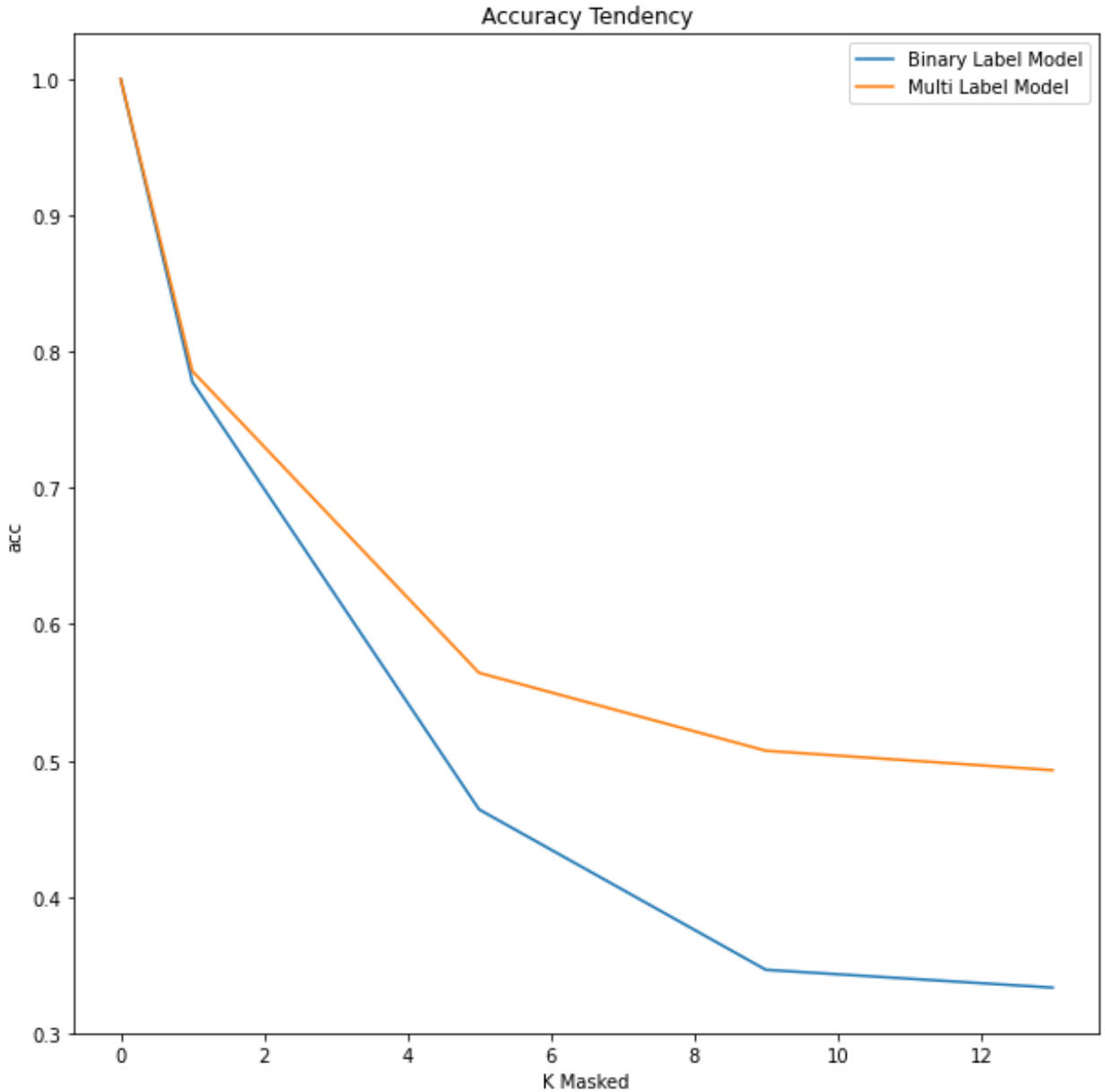
'Right-left disorientation in [UNK] of the [UNK] type. [UNK] [UNK] [UNK] [UNK] -left orientation [UNK] R/L-O) on a confronting subject is more impaired in [UNK] [UNK] dementia [UNK] the Alzheimer type than [UNK] [UNK] [UNK] [UNK] [UNK] infarct [UNK] [UNK] comparable degree of [UNK] . The impairment in R/L-O is independent of aphasia and spatial disorientation.'

Figure: Top K masked

Repeat the same process to all sample on the test set, to see if the accuracy will decrease.

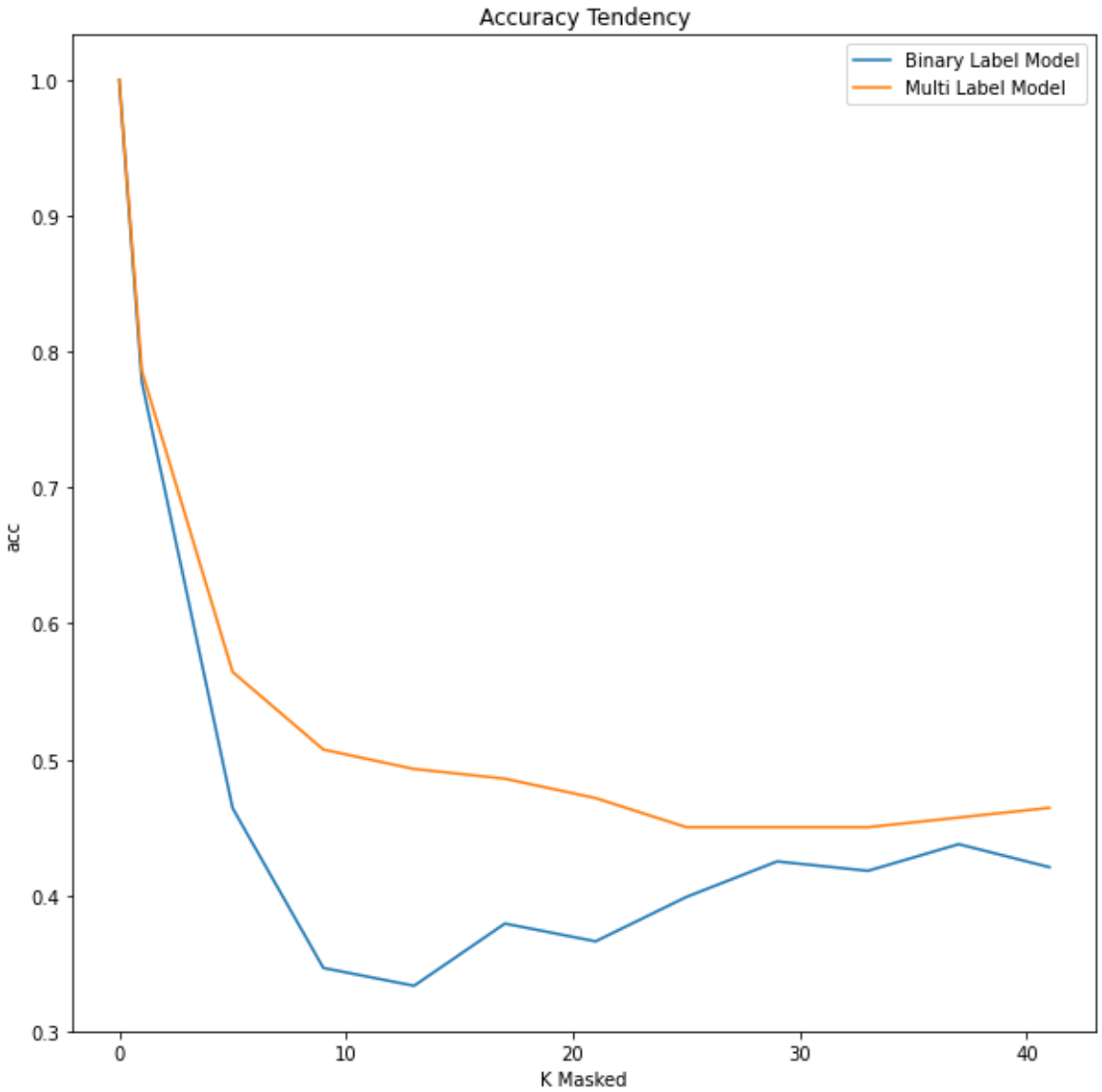
Result

Accuracy Tendency



Result

Accuracy Tendency



Code

Github Link:

$https : /github.com/RmmLeo/STAT6289_Homework/tree/main/Midterm\%20Project$