

# AI Project Report



**Farhan Khalid Butt**

17i-0242

09.12.2020

## Introduction

This project is based on the [SemEval 2020 Task 4 \(ComVE\) Task A](#). Recently utilizing natural language understanding systems to common sense has received significant attention in the research area. Nevertheless, this kind of task remains pretty complicated to solve because of the ambiguity which is one of the natural properties of language and yet to achieve performance compared with human understanding of common sense.

## Data

The data was provided by SemEval to train and test. The task is to predict the sentence which is against common sense out of a pair of sentences. For example,

- He put a turkey into the fridge.
- He put an elephant into the fridge. (Model predicts this as against)

There are a total of 10,000 pairs of sentences available for training, 997 pairs of sentences for validation and 1000 pairs of sentences for testing.

There are 9427 unique words in the training data. Most frequently used words are:

```
the 8719
. 7333
a 6437
to 5650
in 3593
is 3501
I 2295
The 2136
He 2041
on 1753
```

There are 39 pos-tags in the training data. Most frequently used are:

```
NN 32585
DT 19658
IN 15677
```

```
PRP 11348
NNS 9134
JJ 7848
VBD 7840
VB 7829
. 7378
VBZ 6396
```

The original data is in the following format:

**Id sentence0 sentence1**

And labels are in the following format:

**Id label**

What I have done is split each row into two sentences and the labels are assigned as 0 for wrong and 1 for correct. For example,

ID	Sentence	Labels
0	He poured orange juice on his cereal.	0
0	He poured milk on his cereal.	1
2	Jeff ran a mile today	1
2	Jeff ran 100,000 miles today	0

We have labels in the form:

0	0
2	1

This means in ID 0 sentence0 is wrong and in ID 2 sentence1 is wrong. First sentence0 gets labeled and then sentence1 gets labeled, according to this. I have done this for all of the data i.e training, validation and test.

## Methods

### OpenAI GPT

The GPT model is a unidirectional pre-trained model based on probability of the next word in a sequence. We use GPT to look into the perplexity scores. The perplexity score of the sentence means how this sentence doesn't make any sense in some ways. The higher the perplexity, the less sense it makes.

## **Bert:**

Bert is a bidirectional pre-trained model and uses bidirectional encoder representations. I have used the '`bert-large-uncased`' model for bert. We will use train and dev(val) dataset to adjust hyper-parameters and get the best model. I used 32 batch size, 8 training epochs,  $2e-5$  learning rate and  $1e-8$  eps value.

## **DistilBert:**

For DistilBert I have used '`distilbert-base-uncased`'. The batch size for DistilBert 8, 4 training epochs and learning rate  $5e-5$ .

## **Majority Voting:**

As I'm getting predictions from 3 different models, I only take labels if they are present in all of the predicted labels or at least in 2 of them.

## **Results**

### **GPT**

We get an accuracy of around 75% for test data using the GPT model.

### **Bert**

The Bert large-uncased gives us an accuracy of about 98% for training data, 83% for our validation data and 84.5% for testing data.

### **DistilBert**

The Bert large-uncased gives us an accuracy of about 97% for training data, 78.2% for our validation data and 84.9% for testing data.

### **Combined**

The accuracy after doing majority voting is 87.3% on the test set.

## **References**

- <https://huggingface.co/>

- <https://swatimeena989.medium.com/bert-text-classification-using-keras-903671e0207d>