Truth or DeGPTion: Evaluating Lie Detection Capabilities of GPT-3.5 through Fine-Tuning on Personal Opinions, Autobiographical Memories, and Intentions

Tanner Graves
tanneraaron.graves@studenti.unipd.it

Marco Uderzo
marco.uderzo@studenti.unipd.it

Francesco Vo
francesco.vo@studenti.unipd.it

Mehran Faraji
mehran.faraji@studenti.unipd.it

Claudio Palmeri
claudio.palmeri@studenti.unipd.it

Abstract

This paper aims at evaluating the capabilities of GPT3.5 in the task of Lie Detection. This is done through the fine-tuning of GPT3 on three English-language datasets encompassing personal opinions, autobiographical memories, and future intentions. Fine-tuning of LLMs consists in adapting a pre-trained language model to a specific task by further training the model on task-specific data, thereby enhancing its ability to generate contextually relevant and coherent text in line with the desired task objectives. In our investigation, the objective is to discern and classify instances of truth or deception.

1. Introduction

Multiple papers consistently show that the capability of humans to discern truth from deception is at chance level, there is a growing interest in employing Machine Learning methods, especially based on the Transformer Model, to more accurately predict the truthfullness of a statement. Indeed, the inherent pattern recognition capability of ML Models allows them to pick up subtle cues that humans just seem to miss. In this paper, we will use OpenAI's GPT-3.5 Large Language Model (LLM), performing benchmarks on the performance of the base model, and then on a GPT-3.5 model specifically fine-tuned on the Opinion Dataset (Deceptive Opinions), Memory Dataset (Hippocorpus) and Intention Dataset.

2. Methods

2.1. Dataset Preprocessing

(Explain what has been done in the datasets notebook).

2.2. Experimental Setup: Scenarios

We first aggregated the three train and test sets from Scenario 1 (explain what that is). Then we fine-tuned and tested the model on those aggregated sets. This Scenario assesses the capacity of the model to learn and generalize from samples of truthful and deceptive narratives from multiple contexts.

2.3. GPT-3.5 Fine-Tuning

The datasets were formatted into JSON to align with the expected input format of the OpenAI API, subsequently divided into training, validation, and test sets. To manage the potential high computational costs, the model was trained on a subset of the dataset. The model was trained utilizing the OpenAI API, and its performance was assessed through testing and comparison with GPT-3.5. Further experimentation was conducted to assess the impact of engineering the system prompt on overall performance.

Specifically, we noticed that the baseline GPT-3.5 prefers giving verbose or indecisive answers. Verbose answers, that actually classify a statement as genuine or deceptive can be classified easily. Nonetheless, the model decides not to give a definitive answer when it thinks it does not have enough information to classify the statement. The following example shows this behaviour.

```
User: "Each and every abortion
is essentially a tragedy. The
potential mother will suffer
unforeseen consequences. Society
as a whole will be deprived of
the potential it could have
received from the new life."
```

```
Baseline GPT-3.5: "There is no
objective truth to the statement
as it expresses subjective opinions
and beliefs about abortion. It cannot
be definitively classified as
'True' or 'False'."
```

To address this issue it was necessary to engineer a system prompt that discourages this behaviour and adequately explains the task. This prompt is provided to the model at every example query, so instructions should be concise to minimize any token overhead that leads to increased cost of training and queries.

```
System Prompt to Fine-Tuned GPT-3.5:

"You are an expert capable of
discerning truthful from deceptive
opinions based on speech patterns.
Definitively classify the following
statement as 'True' or 'False',
based on the likelihood the statement
represents a genuinely held belief
or a deception."
```

This issue is avoided in the fine-tuned models as the training process rewards our expected behaviour and output format.

3. Results

4. Discussion

5. Code Availability

The datasets used and all the code used for this project is available at the following GitHub Repository.

6. References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [1]. Where appropriate, include the name(s) of editors of referenced books.

References

[1] Authors. The frobnicatable foo filter, 2014. Face and Gesture submission ID 324. Supplied as additional material fg324.pdf.