UNIVERSITY OF PADUA
UNIVERSITA' DEGLI STUDI DI PADOVA

Truth or DeGPTion:

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

Cognitive, Behavioural and Social Data, A.Y. 2023/24

Marco Uderzo, 2096998
Tanner Graves, 2073559
Claudio Palmeri, 2062671
Francesco Vo, 2079413
Mehran Faraji, 2071980

DIPARTIMENTO MATEMATICA

Introduction: Lie Detection

- **Lie-detection** is the task of classifying statements made by humans as either true or false
- Human performance in such task is at **chance level**, whereas computational approaches have been proved to **outperform humans**.
- Our objective is to assess the lie detection capabilities of a popular large language model: **GPT-3.5**.
- We have at our disposal 3 datasets of English sentences regarding different topic:
  - personal opinions
  - autobiographical memories
  - future intentions
- already labelled as true or deceptive statements.
- This data will be used to both to fine-tune our LLM and to test its accuracy according to **Scenario 3** of Loconte et al. paper.

Personal Opinions Dataset

- The participants of this study were divided in 4 groups (HIT's) and asked to provide either a **truthful** or a **deceptive** opinion on the following topics:
  - Abortion
  - Cannabis Legalization
  - Euthanasia
  - Gay Marriage
  - Policy on Migrants

Domain	HIT1	HIT2	HIT3	HIT4
Abo	D	T	D	T
CL	T	T	D	D
Eut	T	D	T	D
GM	T	D	T	D
PoM	D	T	D	T

- This data was extracted both from Italian and English-speaking subjects.
- After removing the entries that didn't respect the instruction given to them (e.g. unintelligible or unreasonably short or descriptive opinions) **2500 entries for each language were gathered in this dataset**.

Autobiographical Memories Dataset

- **6,854** diary-like short stories about **salient life events** gathered in 3 steps from 2 groups of people (A,B)

- **Stage 1: Group A** writes 15-25 sentence **truthful** stories with a 2-3 sentence summary and a timeframe
- **Stage 2: Group B** is tasked to write an **imaginative** story with the **summary of group A** as a prompt
- **Stage 3:** After 2 months **group A** is asked to **retell the story** starting with their summary as a prompt

	# stories	# sents	# words
recalled	2,779	17.8	308.9
imagined	2,756	17.5**	274.2**
retold	1,319	17.3*	296.8**
total	6,854		

Future Intentions Dataset

- All participants are divided into either the **truthful or deceitful** group.
- Participants of the **Truthful Group** were asked to **describe** a non-work-related activity that they'll be doing in the next seven days **answering** the following questions:
  - Q1: "Please describe your activity as **specific** as possible"
  - Q2: "Which information can you give us to **reassure** us that you are telling the **truth**"
- The participants of the **Deceitful Group** were given **3 activities from the former group**, asked which ones **didn't apply** to them and get **randomly assigned** one of those.
- Finally, they were required to **answer Q1,Q2.**
- **1640 examples**

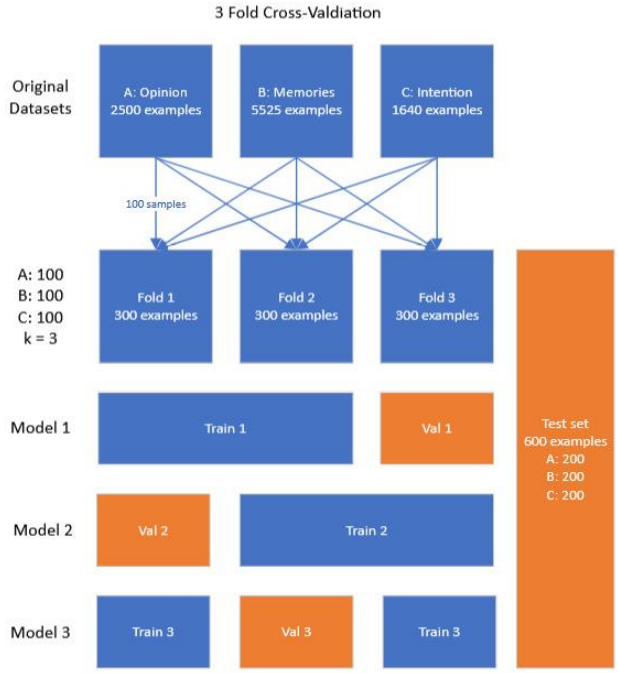
Veracity	Activity	Statement given by participant
Truthful	Going swimming with my daughter	We go to a Waterbabies class every week, where my 16-month-old is learning to swim. We do lots of activities in the water, such as learning to blow bubbles, using floats to aid swimming, splashing and learning how to save themselves should they ever fall in. I find this activity important as I enjoy spending time with my daughter and swimming is an important life skill.
Deceptive	Going swimming with my daughter (assigned)	I will be taking my 8-year-old daughter swimming this Saturday. We'll be going early in the morning, as it's generally a lot quieter at that time, and my daughter is always up early watching cartoons anyway (5 am!). I'm trying to teach her how to swim in the deep end before she starts her new school in September as they have swimming lessons there twice a week.

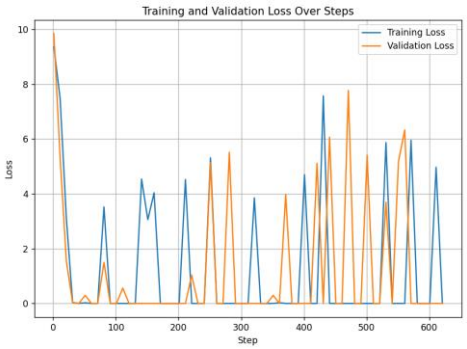
Towards «Scenario 3»

- We have been assigned **“Scenario 3”** from Loconte et al.
- This consist in **aggregating** the aforementioned 3 datasets **into one** in order to create a model capable of detecting truthful statements **in all those scenarios**.
- After having randomly **shuffled** the dataset we selected a training and a test set so that we’d be able to measure the performance of our model.
- In the **Personal Opinions Dataset**, only **English**-language opinions were considered
  - Massively **unbalanced** datasets are problematic and **hinder performance**
- In the **Autobiographical Memories Dataset**, the **retold** category was discarded
- We trained our model on two versions of the dataset:
  - a **subset** of the original dataset (train: 210, validation: 45, test: 45)
  - the **full** dataset (train: 6765, validation: 1450, test: 1450)
- We did this to evaluate the model with different sized datasets.

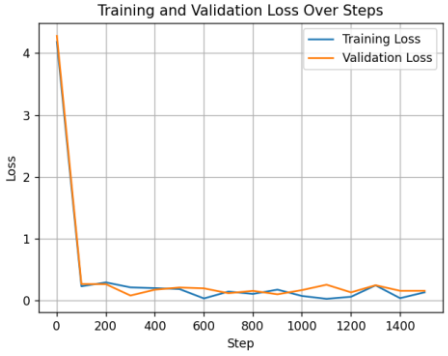
Fine-Tuning GPT-3.5

- We trained our GPT-3.5 model with the following prompt:
  - “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”
- The second sentence is necessary since otherwise the model will often give verbose or indecisive answers.
- Models considered:
  - Baseline GPT-3.5
  - 300-Model
  - Full-Model
  - CV-Model





300-Model Train/Val Losses



Full-Model Train/Val Losses

Results

- Baseline GPT-3.5 performed at chance level.
- 300-Model improved by 16% from baseline.
- Full-Model improved by 32.7% from baseline.
- Cross-Validated Model Improved 9.7% from 300-Model and 25.7% from baseline.

- Interestingly, Future Intentions are the most problematic with the highest variability in performance.

Model	Overrall Accuracy
Baseline GPT-3.5	49.5%
300-Model	65.5%
Full-Model	82.2%
CV-Model	75.2%

Overall Accuracy of the models compared.

Model	Dataset	Accuracy
Base GPT-3.5	Personal Opinions	49.5%
	Autobio. Memories	47.5%
	Future Intentions	51.5%
300-Model	Personal Opinions	79.1%
	Autobio. Memories	67.3%
	Future Intentions	50.2%
Full-Model	Personal Opinions	86.3%
	Autobio. Memories	82.3%
	Future Intentions	69.7%

Class-wise accuracy of the models compared.

	Model 1	Model 2	Model 3
Personal Op.	83.5%	83.5%	80.0%
Autobio. Memories	68.0%	69.5%	70.5%
Future Intentions	72.5%	72.5%	70.0%
Macro avg. Acc.	74.7%	75.2%	73.5%

Class-wise accuracy of Cross-Validated model.

Conclusions

- Even though the baseline model matches human capabilities (which are at the 50/50 chance level), after fine-tuning the model outperforms humans by a large margin.
- GPT-3.5 and LLMs in general are useful tools in Lie Detection tasks.
- There is a positive correlation between size of dataset and performance
- Cross-Validation seems to improve performance on underperforming classes.