Lie detection

Scenario 1 Opinioni

Alejandra Olivia Cruces Andrews

Mario Alejandro Tapia Montero

Chelsie Adelle Renessa Romain

Rovena Llapushi

Agenda

Overview

- Scenario 1 Opinioni: paper summary and results
- GPT-n overview and architecture

Model & interpretation

- GPT 3.5 Fine – tuning results
- Paper results vs. our results
- Conclusions

Paper summary

Goal

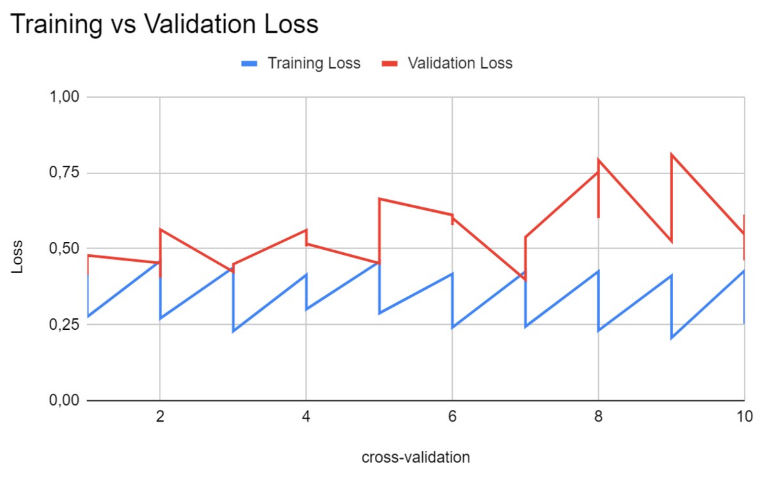
- Determine if the large language model has learned to differentiate linguistic patterns to be deemed a reliable lie detector
- To assess classification performances across 5 different topics

Setup

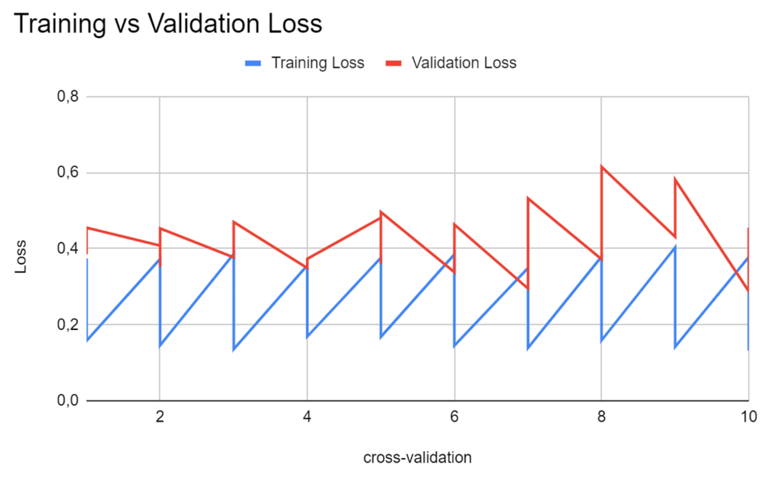
- Fine tune the model on a portion of a **single** dataset and test it on the remaining part of the dataset
- Each opinion was treated as a separate sample
- Training and test sets (no overlap) avoided model bias

Paper results

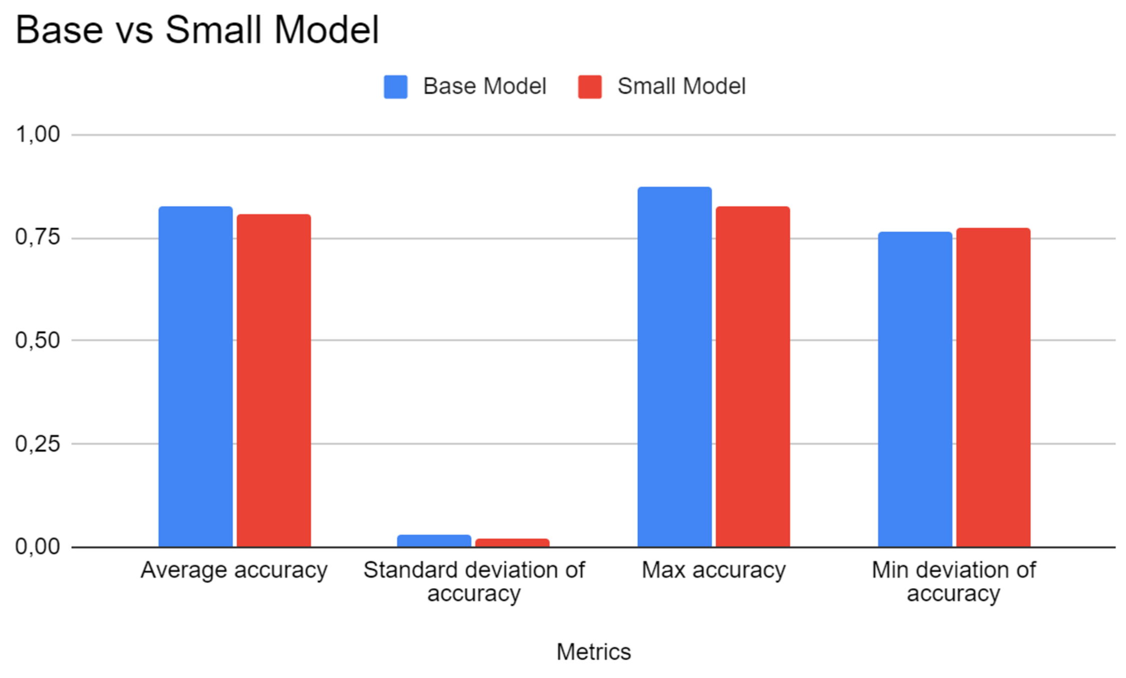
Base model of Flan T5



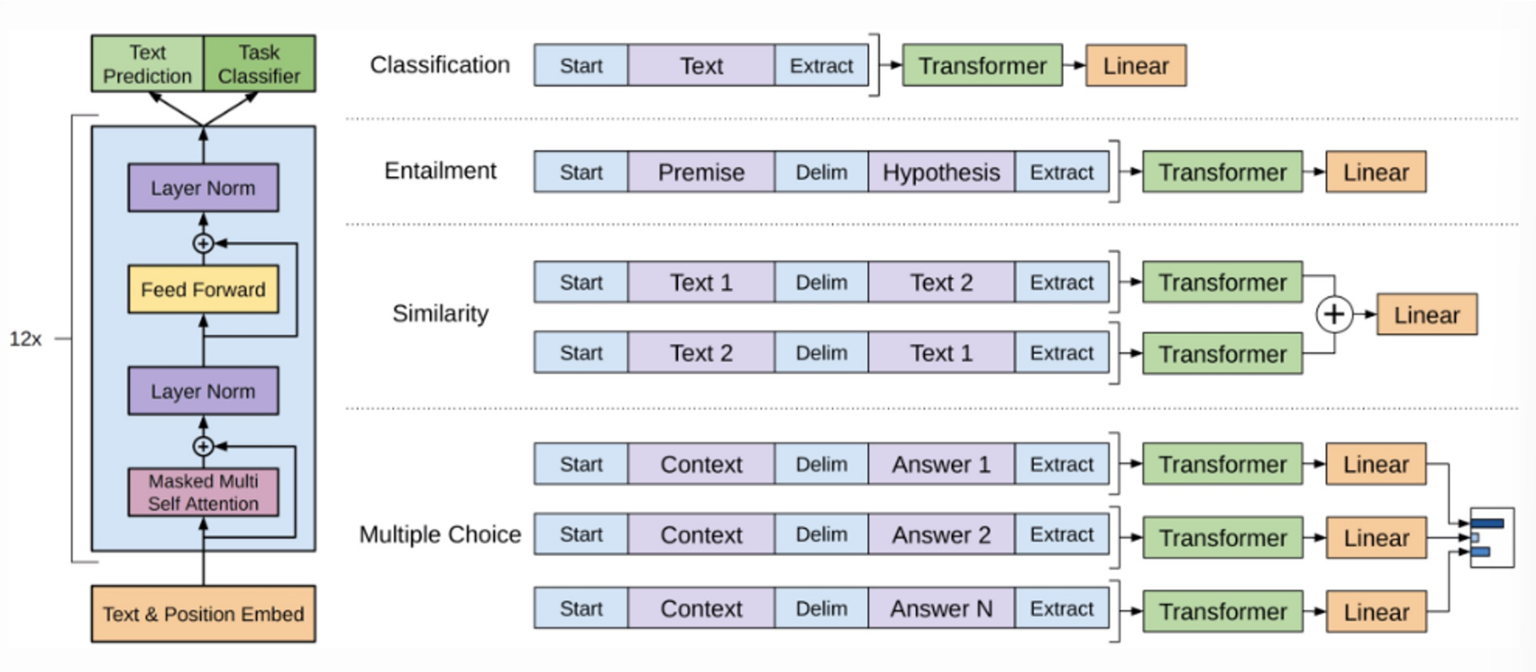
Small model of Flan T5



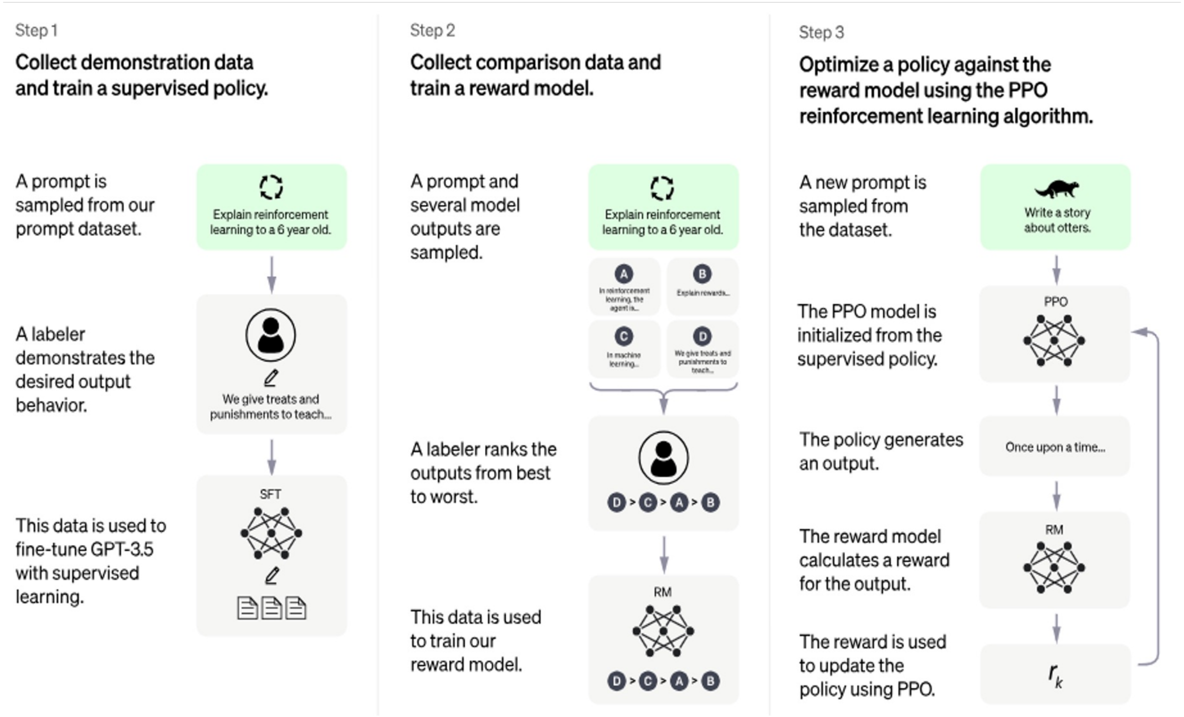
Paper results



GPT-n fine-tuning



GPT-3.5 architecture



https://openai.com/blog/chatgpt

GPT-3.5 model setup

Setup

- In order to use `GPT-3.5-turbo-0163`, the API requires that the following are specified:
  - System
  - User
  - Assistant

- System role– "Given a statement, your task is to determine if the user is **honest** or not about their opinion."
- User – row sent (the opinion)
- Assistant – row label, True (T) or False (F)

GPT-3.5 model setup (continued)

Setup

- The latter is specified as a .jsonl file with the following format

```
{"message": [{'role': 'system', 'content': System_role},
{'role': 'user', 'content': Opinion},
{'role': 'assistant', 'content': Labels}, …]}
```

GPT-3.5 model setup (continued)

Setup

- Dataset contains 2500 examples
- 4 cross validation folds were used with a **75%-25%** train-validation split
- This achieved a balance between budget constraints and model performance
- 3 epochs were used for training
- This is the same number of epochs used in the paper, to achieve the maximum test accuracy without overfitting

GPT-3.5 fine-tuning procedure

Steps

1. We fine-tune the model with the training data set

1. We test the fine-tuned model on the validation set

1. We calculate the metrics: accuracy, precision, recall, F-score

GPT 3.5 results

Summary of the metrics for evaluating the models by CV fold

	Accuracy	Precision	Recall	F-score
Split 1	0.8816	0.8692	0.8971	0.8829
Split 2	0.8512	0.8904	0.8100	0.8483
Split 3	0.8688	0.8775	0.8548	0.8660
Split 4	0.8624	0.8558	0.8669	0.8571
Average value	0.8660	0.8732	0.8572	0.8636
Standard deviation	0.0110	0.0126	0.0313	0.0128

- High and reasonable accuracy for an LLM
- Models perform similar, indicating the absence of overfitting

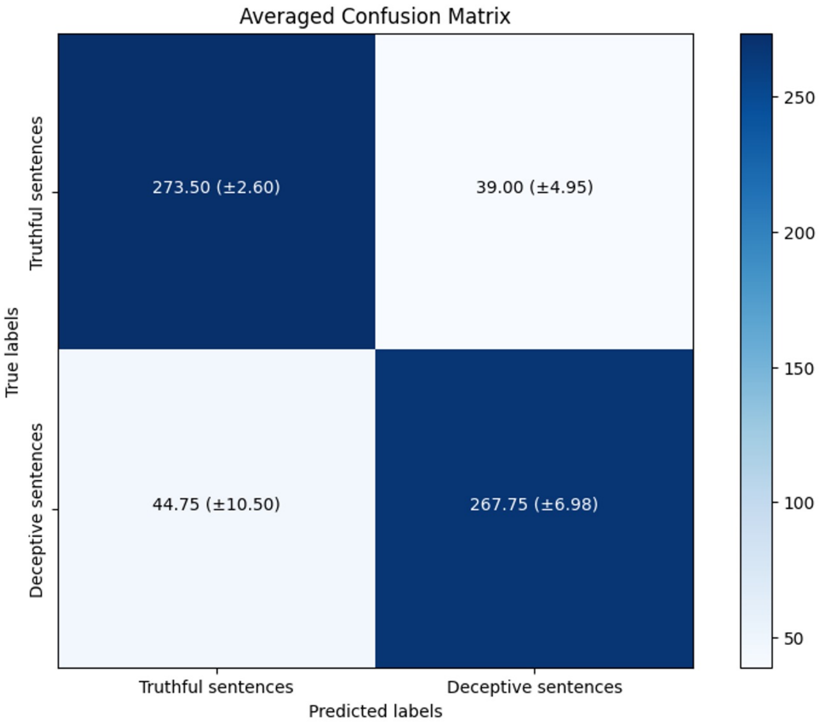
GPT-3.5 results (continued)

Summary of the average metrics for evaluating the models by topic

	Accuracy	Precision	Recall	F-score
Abortion	0.8520	0.8788	0.8164	0.8460
Euthanasia	0.8680	0.8778	0.8579	0.8660
Immigration	0.8820	0.8935	0.8676	0.8803
Gay Marriage	**0.8900**	0.8943	0.8870	0.8893
Cannabis Legalization	**0.8380**	0.8254	0.8567	0.8400
Standard deviation	0.0213	0.0283	0.0258	0.0212

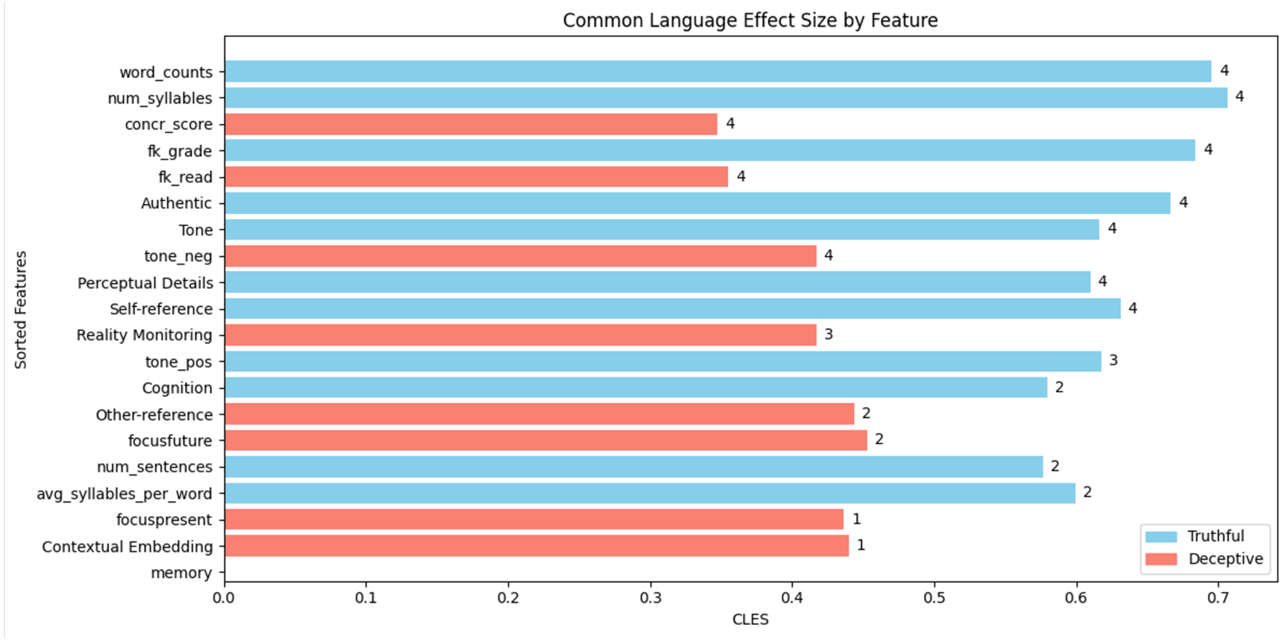
- Evidence that the classifier is more effective on topics that are more polarized and less ambiguous, such as Gay Marriage (best performance) than on topics that are more nuanced and complex, such as Cannabis Legalization (worst performance)

GPT-3.5 results (continued)



- Error classification: Higher number of FP compare to FN
- GPT 3.5 was more likely to classify a result as true, regardless of whether the opinion was truthful or not

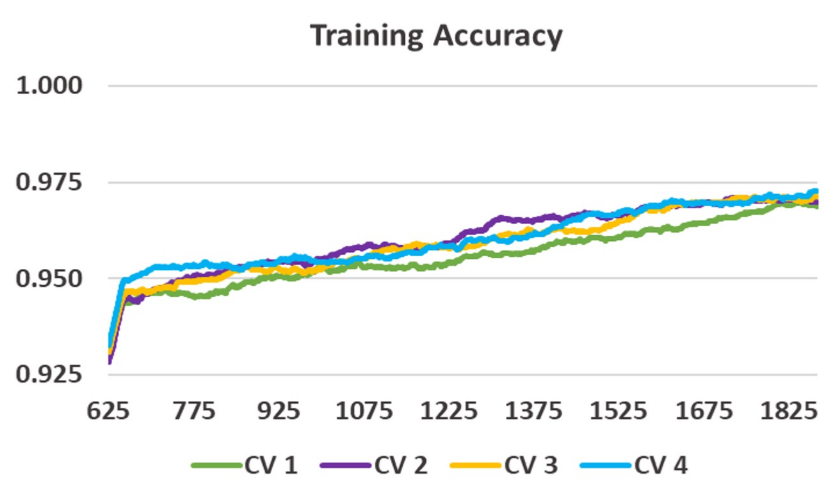
GPT-3.5 results (continued)

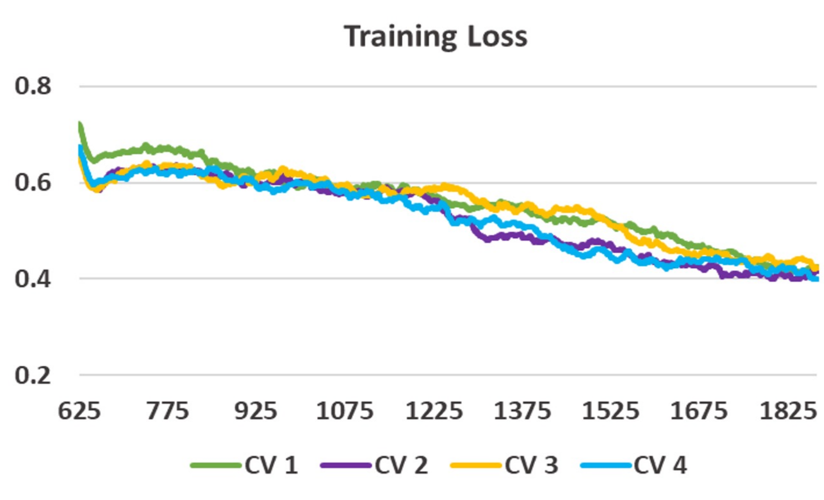


CLES: Probability that a specific linguistic feature, in a picked-at-random truthful statement, will have a higher score than in a picked-at-random deceptive one.

GPT-3.5 results (continued)

Moving average of 625 steps (1 epoch) training accuracy and training loss for each CV fold





GPT-3.5 results (continued)

Comparison with FLAN T5 on test set

Model	Average accuracy	Standard deviation
Flan-T5 small	0.8064	0.02
Flan-T5 base	0.8260	0.03
GPT-3.5-turbo-0163	**0.8660**	0.0110

- GPT3.5 Turbo has better generalization capabilities than FLAN T5 models

Applications

- Possible use in Law Enforcement and Forensics, where accuracy of detecting deception is crucial
- If model produces too many false positives -> many liars are misclassified as truthful
- If model produces too many false negatives -> truth-tellers are misclassified as liars
- F-score is used to balance between precision and recall
- A value of 0.8660 demonstrates the robustness and reliability of the model for assisting in these fields

Conclusions

- We have found that the average accuracy (86.6%) of GPT 3.5 on the task of Lie Detection on opinions is higher than the average accuracy (82.6%) of FLAN-T5 base on the same task.

- The top three most significant linguistic features across all splits of the data are:

    1. Number of syllables (Truthful)
    2. Word Count (Truthful)
    3. Concreteness score (Deceptive)

Thanks!