Lie Detection - Opinions

Cognitive, Behavioral and Social Data (2023/2024)

Agata Garbin
agata.garbin@studenti.unipd.it
2072693

Alessia d'Addario
alessia.daddario@studenti.unipd.it
2086506

Alice Ronzoni
alice.ronzoni@studenti.unipd.it
2076675

Graziana Capurso
graziana.capurso@studenti.unipd.it
2097099

In recent years, Large Language Models have been increasingly used in a vast range of problems, from text analysis to sentiment recognition. In this context, we want to focus on the application of LLMs to Lie Detection, with a specific emphasis on opinions.

1. Introduction

Building upon the work of Loconte et.al [1], the aim of this paper is to explore BART capability in understanding whether an opinion is truthful or not, while investigating the difference between our chosen model and FLAN-T5.
The dataset utilized for this task comprises 2500 opinions provided by 500 different individuals regarding 5 different topics evenly split between truthful and deceptive.
Furthermore, we extended our analysis to include the Memories and Intention datasets.

2. Lie Detection

Lie detection involves the process of determining whether a given communication is true or false. Research has shown[2] that, when telling a lie, a subject is prone to using verbal strategies to induce false beliefs in the interlocutor, leading to a specific temporary psychological and emotional state.

Starting from this behavioral assumption, the Undeutsch hypothesis suggests that deceptive narratives diverge both in form and content from truthful narratives[3]. For this reason, it is not absurd to consider the possibility of defining ad hoc strategies in the field of lie detection with potential applications in forensic and legal settings.
Numerous studies have focused on identifying verbal cues to differentiate between truthful and deceptive narratives, although it can be argued

that human performance in correctly utilizing these verbal cues is not satisfactory. This phenomenon may be explained by the ”truth bias”[4], i.e., the presumption of the partner's honesty.

Given this innate human inability, studies have employed computational techniques such as stylometry, which involves using computational linguistic and artificial intelligence tools to quantitatively analyze written texts. This helps uncover unique patterns that can indicate authorship or other stylistic traits.
In this regard, many recent studies have been conducted using Machine Learning and Deep Learning algorithms combined with Natural Language Processing (NLP) techniques to automatically detect deception from verbal cues. Particularly, among various possible implementations of NLP, satisfactory results have been obtained by using fine-tuned Large Language Models (LLMs) on small corpora for lie detection tasks.

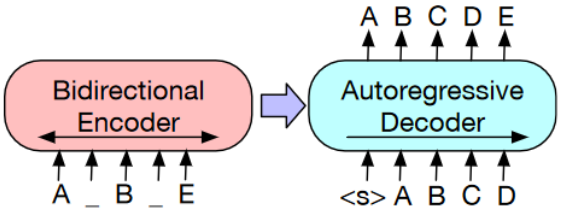
As already mentioned, the starting point of this paper is the work of Loconte et al.[1], a study of a fine-tuned open-source LLM named FLAN-T5, developed by Google, used in the lie detection problem using three datasets encompassing personal opinions (the Deceptive Opinions dataset,[5]), autobiographical experiences (the Hippocorpus dataset,[6]), and future intentions (the Intention dataset,[7]). In particular, we focus on the Opinions Dataset described in[5], i.e., DecOp (Deceptive Opinions), a new multilingual and multi-domain corpus for the automatic detection of deception in typed text, which will be further described in the Dataset Section.

3. Methods

To ensure comparable results, given that the FLAN-T5 model has a sequence-to-sequence architecture, we opted for a model of the same type: BART.

3.1. BART

BART is a denoising autoencoder designed for a wide range of tasks, employing a sequence-to-sequence model architecture. [6] It enhances upon BERT's objectives by focusing on longer-range transformations and overall sentence structure understanding. BART is applicable for pre-training and fine-tuning stages, facilitating tasks such as sequence classification, token classification and machine translation.
BART utilizes a standard sequence-to-sequence Transformer architecture, with ReLU activations replaced by GeLUs. Parameters are initialized from a normal distribution with mean 0 and standard deviation 0.02. The base model is built with 6 layers in encoder and decoder, while the large model is built with 12 layers each, containing approximately 10% more parameters than equivalent BERT models. The architecture is closely related to that used in BERT, with the following differences: (1) each layer of the decoder additionally performs cross-attention over the final hidden layer of the encoder; and (2) BERT uses an additional feed-forward network before word- prediction. BART combines aspects of both BERT and GPT, using a denoising autoencoder pre-training approach.



BART is a versatile and powerful tool for natural language processing, thanks to its ability to tackle different sequence-to-sequence tasks effectively. From summarizing documents to translating languages, BART proves to be valuable across various applications. Its strong architecture and well-designed training techniques play key roles in its success during both pre-training and fine-tuning phases.

3.2. Dataset description

The dataset used in our analysis was first presented in the paper " A multilingual and multidomain corpus for detecting deception in typed tex" [5]. The authors of this paper introduce a new multilingual and multi-domain dataset called DecOp (Deceptive Opinions) aimed at automatically identifying deception in written language, with the goal of enhancing understanding of automatic deception detection methods. The DecOp corpus includes personal perspectives across five different domains that are both real and false. It is carefully designed to support different kinds of categorization experiments and is available in two languages.
The dataset contains the opinions from 500 people on five distinct topics: abortion, cannabis legalization, euthanasia, gay marriage and policy on migrants, resulting in a total of 2500 sentences. In particular, it consists in a total of 1250 truthful and 1250 deceptive opinions balanced across topics. The domain selection process starts from the premise that most individuals are likely to have an opinion on these subjects and can readily express or disagree with it. The ground truth for the applied paradigm is an experimental one. This indicates that each respondent gave, in a free text answer modality, their honest and deceptive first-person opinions on the mentioned subject in accordance with predetermined guidelines, in at least 4 or 5 lines. To preserve the balance between honest and untrue ideas, four Human Insights Assignments (HITs) were made and adjusted for ground-truth. The test set has been uniformly sampled on the four HITs. Consequently, the distribution of labels is maintained.

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

Participants were both from United States and Italy and, to facilitate cross-linguistic comparisons, they responded in their native languages. In the final version of the DecOp each opinion is labelled with its domain, veracity condition (Truthful or Deceptive), and language (standard American English or Italian). For each participant age and gender are also given.

TRUTHFUL	DECEPTIVE
DOMAIN: ABORTION	
IT	
Penso che ogni donna dovrebbe avere diritto di scegliere se portare avanti o meno una gravidanza. In ogni caso non mi schiero completamente a favore dell'aborto, perchè ci sono circostanze in cui viene comunque interrotta una gravidanza nonostante potrebbero esserci soluzioni alternative, che non comportino né un cambiamento di vita drastico per la donna, né la perdita di una vita.	L'aborto è una cosa inumana e non capisco come possa essere legale. Fortunatamente esistono gli obiettori di coscienza che decidono al posto di quelle sciagurate che hanno pensato bene di rimanere incinta e poi se ne pentono e vogliono uccidere il bambino. Che poi, cosa costa portare a termine la gravidanza e dare in adozione il bambino?
EN	
While I am morally torn on the issue, I believe that ultimately it is a woman's body and she should be able to do with it as she pleases. I belive people should not dehumanize the fetus tough, to make themselves feel better. The decision about laws regarding this issue should be left up to the states to decide. To combat this problem, birth control should be easily accessible.	Abortion is the termination of a life and should not be allowed. If a fetus has made it to the point of being able to survive "on its own" outside its mother's body, what right do we have to cut its life short. If the mother's life is in danger, she already chose that she was willing to sacrifice her life to have a child when she consented to procreating.

Table 3: The table shows some examples of the gathered opinions for both the Italian (IT) and standard American English (EN) languages included in the DecOp corpus. The domain considered for the example is Abortion.

3.3. Fine Tuning

Fine tuning in machine learning is the process of adapting a pre-trained model for specific tasks or use cases. [8]
A Large Language Model (LLM) that has been trained on a large dataset possess a vast knowledge but may not perform well on a specific domain, for this reason, in order to make the LLM more accurate on the new dataset, fine-tuning can be used.

Fine tuning is a technique very used in Deep Learning and especially in LLM because, instead of training from scratch a model, we can hone the capabilities of a pre-trained model and adapt them to solve the new task. Essentially, fine-tuning involves leveraging the parameter weights of the pre-trained model to address the new problem, this results in the reduction of computing power and labeled data used to obtain a valuable result.

For our purpose we imported the pre-trained weights from the Hugging Face library `https://huggingface.co/facebook/bart-base` where BART has been trained on corrupted documents.

Following Loconte's directions, 450 writers have been used to fine tune the model and the remaining 50 to test it. In this context we refer to "writers" and not "sentences" because in order to avoid the model to infer and recognize the style of a single individual, we made sure that opinions provided by the same person were considered in the same set.

4. Results

In this section we report the most significant results obtained from multiple fine tuning experiments that were held using different combinations of parameters.
Given that there are two variations of BART model – base and large – for thoroughness, we conducted experiments with both variants.

4.1. BART Base

As baseline, here we show the results of the model using default parameters:

cross validation	10
epochs	3
learning rate	5e-5
batch size	2
weight decay	0.01
accuracy	0.806 ± 0.0148

And here is the configuration that yielded the best results for the base model:

cross validation	4
epochs	3
learning rate	5e-5
batch size	20
weight decay	0.01
accuracy	0.8224 ± 0.0117

To be exhaustive, here we present some of the other experiments that we conducted, with the first listed being the least successful among all:

- 1

cross validation	5
epochs	3
learning rate	5e-6
batch size	20
weight decay	0.01
accuracy	0.7636 ± 0.0182

- 2

cross validation	5
epochs	5
learning rate	5e-6
batch size	2
weight decay	0.1
accuracy	0.82 ± 0.0140

- 3

cross validation	5
epochs	3
learning rate	5e-6
batch size	10
weight decay	0.1
accuracy	0.7812 ± 0.0159

- 4

cross validation	5
epochs	3
learning rate	5e-5
batch size	20
weight decay	0.1
accuracy	0.8124 ± 0.0047

Once obtained the best results after fine-tuning on the Opinion dataset, we decided to do some trials also using the Memories[6] and the Intentions[7] datasets.
Overall, we observed that the average accuracy was lower than the one from our original experiments.

The following are the best sets of parameters found:

- Intention

cross validation	5
epochs	5
learning rate	5e-6
batch size	2
weight decay	0.1
accuracy	0.72561 ± 0.0327

- Memories

cross validation	5
epochs	5
learning rate	5e-6
batch size	2
weight decay	0.01
accuracy	0.794095 ± 0.0165

As we can notice, in both cases, the best performance was achieved by the same configurations of parameters which, however, differ from the one found to be the best for the Opinion dataset.

4.2. BART Large

Here is the best result obtained after fine-tuning BART large using the opinion dataset, the result is quite satisfactory even though the accuracy is lower than the base model.

cross validation	10
epochs	3
learning rate	5e-5
batch size	2
weight decay	0.01
accuracy	0.8112 ± 0.014943

4.3. Comparison

From Loconte's paper[1]:

Model	Opinion	Memory	Intention
Flan-T5 small - Scenario 1	80.64 ± 0.02	76.87 ± 0.02	71.46 ± 0.03
Flan-T5 base - Scenario 1	82.6 ± 0.03	**80.61 ± 0.01**	71.52 ± 0.02
Flan-T5 small - Scenario 3	79 ± 0.02	75.67 ± 0.02	69.32 ± 0.037
Flan-T5 base - Scenario 3	**82.72 ± 0.024**	79.87 ± 0.016	**72.25 ± 0.03**

Our best results:

Model	Opinion	Memory	Intention
BART	0.82 ± 0.01	0.79 ± 0.01	0.72 ± 0.03

For fairness, we consider only a comparison between the base models.

As we can see, the original paper achieved a value of 0.826 on the Opinions dataset, which is not so different from our result. Moreover, as expected, in both cases, the others two datasets produced lower accuracy values.

4.3.1 Comments

The results produced by the BART model are quite promising, especially for the Opinion dataset. This suggests that it is indeed possible to train a model to discern whether a sentence is truthful or not.
If we look at the results produced by the model in more detail, we can observe that increasing the number of iterations doesn't always lead to an increase in accuracy. However, we can also notice that larger batch sizes tend to result in higher precision, but, this remains true only if the learning rate is big enough. In general it would be interesting to increase the number of iterations for each model to see if the average precision increases.
Moreover, comparing BART results' with the ones generated by FLAN-T5 we can argue that the difference is not so large, this may suggest some sort of equivalence between the two models that could be further investigated by expanding the study considering other LLMs.

4.4. Bart on the Italian Dataset

As previously mentioned3.2, the paper [5] not only introduced a new English dataset but also an Italian one, both sharing the same structure.
Eager to see what BART was capable of achieving with the Italian language, we decided to fine tuned the model on this new dataset, using the same setting described above.
In order to do so, we made use of a tailored version of Bart pre-trained on Italian text corpora, provided by Hugging Face `https://huggingface.co/morenolq/bart-it`.
Here we report some experiment, considering the same parameters configuration of 4.1, for the sake of comparison:

- 1

cross validation	4
epochs	3
learning rate	5e-5
batch size	20
weight decay	0.01
accuracy	0.856 ± 0.0126

- 2

cross validation	5
epochs	3
learning rate	5e-6
batch size	20
weight decay	0.01
accuracy	0.793 ± 0.0154

- 3

cross validation	5
epochs	5
learning rate	5e-6
batch size	2
weight decay	0.1
accuracy	0.857 ± 0.0128

As we can see, these results are slightly better than those of the English-based counterpart. This could likely be attributed to the different pre-training processes of the two models.

5. Conclusion and Discussion

The aim of this project was to fine tune a Large Language Model in order to identify if the opinions of 500 different people on 5 different topics (abortion, cannabis legalization, euthanasia, gay marriage and policy on migrants) would be recognized as true or false.
The LLM that we chose to use is BART: a denoising autoencoder employing a sequence-to-sequence architecture.
In order to solve this problem, we fine-tuned the model using various combinations of parameters, resulting in an average precision value of 0.8224.
Comparing this result with the one obtained in the paper "Verbal Lie Detection using Large Language Models"[1] we observe that despite achieving a slightly lower value, the difference between the two models is not substantial. This could be attributed to the structural similarity of the two models and the effectiveness of the fine-tuning technique.
To conclude, we can see that the results are satisfactory, suggesting promising prospects for the use of LLMs such as BART in the field of Lie Detection.
Moreover, the additional experiments involving the Italian language give hope for the potential application of these techniques in languages beyond English.

References

[1] Loconte et al. Verbal lie detection using large language models. 2023.

Ethnic similarities and differences in linguistic indicators of veracity and lying in a moderately high stakes scenario. *Journal of Police and Criminal Psychology*, 30(1):15–26. (2015b).

Cross-cultural deception detection . In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 440–445 (2014).

[2] Walczyk, J. J., Harris, L. L., Duck, T. K., & Mulay, D. A social-cognitive framework for understanding serious lies: Activation-decision-construction-action theory. *New Ideas in Psychology*, **34**, 22–36.https://doi.org/10.1016/j.newideapsych.2014.03.001(2014).

[3] Amado, B. G., Arce, R., & Fariña, F. Undeutsch hypothesis and Criteria Based Content Analysis: A meta-analytic review. *The European Journal of Psychology Applied to Legal Context*,**7**, 3–12. https://doi.org/10.1016/j.ejpal.2014.11.002(2015).

[4] Levine, T. R., Park, H. S., & McCornack, S. A. Accuracy in detecting truths and lies: Documenting the "veracity effect". *Communication Monographs* **66**, 125–144. https://doi.org/10.1080/03637759909376468(1999).

[5] Capuozzo, P., Lauriola, I., Strapparava, C., Aiolli, F., & Sartori, G. DecOp: A multilingual and multi- domain corpus for detecting deception in typed text. In *Proceedings of the 12th Language Resources and Evaluation Conference* 1423-1430, (2020, May).

[6] Mike Lewis and Yinhan Liu and Naman Goyal and Marjan Ghazvininejad and Abdelrahman Mohamed and Omer Levy and Ves Stoyanov and Luke Zettlemoyer, *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension* https://arxiv.org/abs/1910.13461v1 (2019)

[7] Sap, M., Horvitz, E., Choi, Y., Smith, N. A., & Pennebaker, J. Recollection versus imagination: Exploring human memory and cognition via neural language models. In *Proceedings of the 58th annual meeting of the association for computational linguistics* 1970-1978, http://dx.doi.org/10.18653/v1/2020.acl-main.178(2020, July).

[8] Kleinberg, B., & Verschuere, B. How humans impair automated deception detection performance. *Acta Psychologica* **213**, https://doi.org/10.1016/j.actpsy.2020.103250(2021).

[9] Chung, H. W., et al. Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416. (2022). https://arxiv.org/abs/2210.11416