Fingerprinting Fine-tuned Language Models in the Wild

Nirav Diwan, Tanmoy Chakravorty, Zubair Shafiq

Findings of ACL-IJCNLP 2021
&
RepL4NLP 2021 : The 6th Workshop on Representation Learning for NLP

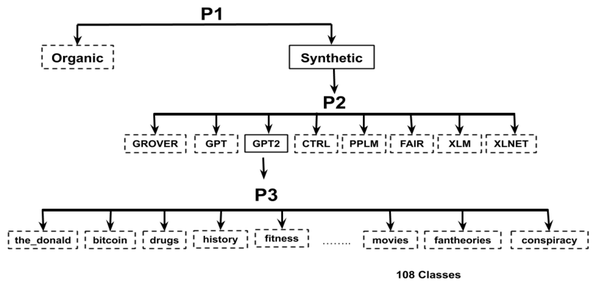
Motivation

The ability of language models (LMs) to generate high quality synthetic text can be misused to launch spam, disinformation, or propaganda. For example, fine-tuned GPT2 LM has been shown generate synthetic text which is more detailed, coherent and grammatically accurate than human written organic text [1]. Fine-tuning enables bad actors to repurpose off-the-shelf pre-trained language models for malicious purposes. Thus, to know the origin of the text, it is important to consider fine-tuned language models.

Problem Statement

authorship attribution of synthetic text generated by fine-tuned language models

Introduction



Origin of text - In the real-world, the problem of origin of text is a multi-level classification problem.

P1- Categorising the text into organic or synthetic.

P2- Attribution of pre-trained LM to synthetic text.

P3- Attribution of fine-tuned LM to synthetic text.

Challenges of P3 - Prior work has largely dealt with P1 and P2 [2,3]. Most notably, there exists no **no real-world data** for the problem and there also a **large number of classes** representing fine-tuned language models.

Dataset

SubSimulatorGPT2 Subreddit

Figure: Each user on the subreddit is a GPT2 bot that is fine-tuned on posts and comments from a particular subreddit, like r/wallstreetbets.

Real-world dataset: Synthetic text available on the r/SUBSIMULATORGPT2 subreddit.

Large number of classes: The subreddit contains posts authored by 108 fine-tuned GPT2 models.

Interaction between bots: Bots also author replies using the synthetic text in the preceding comments/replies as their prompt.

How *good* is the synthetic text?

Figure: Examples upvoted popularly by the subreddit community. The top voted posts are rife with hate speech and impersonation.

Coherency: Synthetic text is coherent although with lower overall readability than organic text.

Characteristics: Synthetic text captures the lexical, vocabulary and readability characteristics of the organic text used to fine-tune it.

Experiments

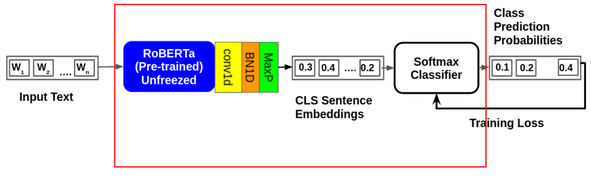
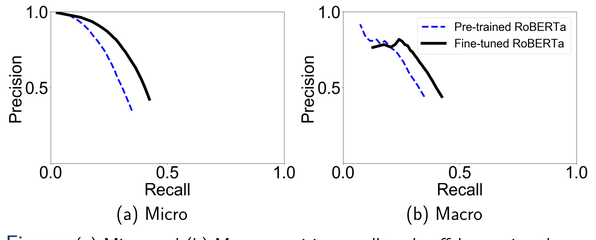


Figure: Fine-tuned RoBERTa followed by a CNN based classifier achieves the best precision and recall.

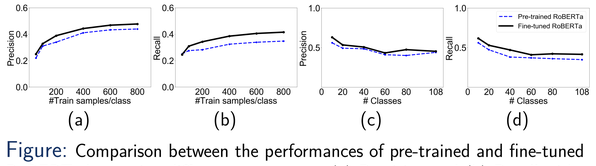
Architecture	Classifier	Macro		Top-k	
		Prec	Recall	5	10
GLTR	RF	7.8	6.6	12.6	19.0
Writeprints	MLP	16.9	14.7	30.8	42.1
GloVE	CNN	31.1	26.7	44.2	53.5
GPT2	MLP	44.9	29.0	47.5	56.9
RoBERTa	MLP	44.0	34.8	54.8	62.5
FT-GPT2	CNN	44.6	42.1	60.9	68.9
FT-RoBERTa	CNN	**46.0**	**43.6**	**62.0**	**69.7**

Table: Performance of multi-class classifiers based on macro Precision (Prec), Recall and top-k accuracy ($k = 5, 10$) for the largest setting of 108 classes.



(a) Micro (b) Macro

Figure: (a) Micro and (b) Macro *precision-recall trade-off* by varying the gap statistic threshold. Micro-precision of 87% is achieved by the best model with only a slight decrease in micro-recall (27%).



(a) (b) (c) (d)

Figure: Comparison between the performances of pre-trained and fine-tuned RoBERTa by varying different parameters. (a) Precision and (b) Recall with the varying training size. (c) Precision and (c) Recall with the varying number of classes. Overall, fine-tuned RoBERTa outperforms pre-trained RoBERTa. The comparison with *all baselines* is included in the appendix.

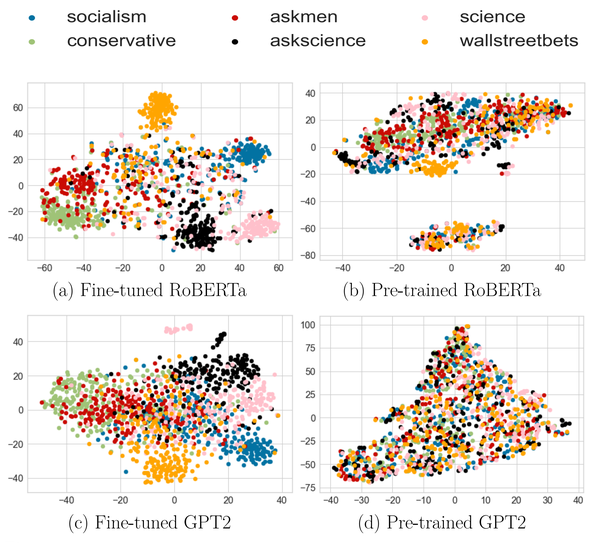
Paper Link -
https://aclanthology.org/2021.findings-acl.409/

Key insights

Stylometric, and other authorship attribution models shown to be accurate for P1 and P2, are not accurate for P3

Fine-tuning itself is the most effective in attributing the synthetic text generated by fine-tuned LMs

Why fine-tuning works well?



(a) Fine-tuned RoBERTa (b) Pre-trained RoBERTa
(c) Fine-tuned GPT2 (d) Pre-trained GPT2

Figure: Visualisation of fine-tuned (a,c) and pre-trained embeddings (b,d) of specific classes. Closely condensed clusters specific to the domain of the organic text form in the fine-tuned embeddings.

References

1. https://www.nytimes.com/interactive/2019/06/07/technology/ai-text-disinformation.html
2. Gehrmann, Sebastian, Hendrik Strobelt, and Alexander M. Rush. "Gltr: Statistical detection and visualization of generated text." (2019).
3. Uchendu, Adaku, et al. "Authorship Attribution for Neural Text Generation." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). (2020).

Contact Information

Have a question? Feel free to reach out!

Nirav Diwan : nirav17072@iiitd.ac.in