CIS 6930: Privacy & Machine Learning
Final Project Report: **Deep learning-based approach for stylometric De-Anonymization for model authorship**

Apoorv Chandurkar
(Point of Contact)
apoorvchandurkar@ufl.edu

Kunwardeep Singh
kunwardeep.singh@ufl.edu

December 12, 2019

1 Introduction

With the rapid advancement in Natural Language Processing and Deep learning, the end-to-end pipeline for many natural language-related tasks is possible. New architectures for Language modeling are now achieving impressive results in Natural Language Generation which can output coherent text given some textual prompt. However, there is a concern that the misuse of these models can be done by spreading fake news and malicious content on a large scale. Hence, from a security perspective, it is imperative to treat these models as potential authors and perform a stylometric analysis of the text they are generating to de-anonymize them to check if they are deployed with malicious intent. In this project, we have developed a Deep-learning based classifier that can classify text generated by the top 3 State of the Art open-source Language Models. Our goal is to research how much effect the architecture of a neural network has on its text generation style if any. We aim to find out if there is something unique about the text being generated by the model and if we can use this property to de-anonymize the text. We add another class of human-generated text and examine how much stylometric difference is present between current SoTA Language Models and human-written text.

2 Background

A statistical language model is a probability distribution over sequences of words. In machine learning problems in the domain of natural language processing, words are embedded into vectors and given as an input to neural networks to perform different tasks like classification, text summarization, question answering, text generation, etc. In the specific case of text generation, language modes are directly relevant to the final output of the neural model as text generated by sampling from the probability distribution of the language model which maximizes the overall probability of the sequence of tokens.

Earlier research used recurrent neural networks and their derivatives like Gated Recurrent Unit or Long-Short Term Memory (LSTM) networks to model a probability distribution over a set of vocabulary. Then, Transformer[1] based architecture was introduced which used the concept of attention, which addressed the issue of sequential processing and limitation of long-term dependency tracking. Hence, there was a significant increase in the SoTA performance on various Natural Language Processing benchmarks. In February 2019, OpenAI released the GPT- 2[2] model paper in which they said that only training a large network of 1.5 Billion parameters to predict the next word given a previous sequence was enough to achieve SoTA performance in text generation task. This was just a scaled-up version of the original transformer-based GPT[3] model. Also, XLNET[4] architecture released which is a large scale transformer-based architecture currently having a top - 5 leaderboard position for the general-purpose NLP task-based GLUE[5] benchmark.

3 Related Work

De-anonymization of anonymous authors through stylometry has been used historically to get literary, historical and criminal investigation breakthroughs. Stylometry was earlier done only manually to identify hidden attributes associated with authors helping in the de-anonymization process. With the advent of machine learning researchers in the security domain have applied machine learning to classify text or attribute authorship based on authorship style or genre. (for eg. Afroz et al. [6], Ramyaa et al. [7]) But in these methods, the features based on which classification is done are hand-crafted and extracted manually. Also, Caliskan-Islam et al [8] used stylometry on programming code to de-anonymize programmers who authored the source code using a Random Forest based classifier. Many papers use stylometry to only predict some personal attributes of the author like gender and/or age etc.(for eg. Sarawgi et al. [9], Surendran et al.[10]). We are proposing to use Deep Learning to classify input text without any completely manual feature extraction method. Also, we are proposing a novel idea of treating AI-models as authors and differentiating between their stylometric style along with researching style differences between humans and Language models.

4 Data Generation

We picked the top 3 State of The Art language models in the literature, namely, OpenAI GPT, GPT2 and XL-NET. After looking and testing different open-source implementation of these models, we decided to use the HuggingFace transformer library (https://github.com/huggingface/transformers) [11] for selecting and generating text. We studied the Transformers library, which has the option to choose and fine-tune many hyperparameters at the text generation phase. We decided on many hyperparameters like no. of tokens generated by the model and input prompt given to the text.

We discussed that to be consistent across different models, input prompt given to generate text should be the same for all models. This will help in defining the classification task as each model will produce text based on the same context in the same domain. Hence, our de-anonymization classifier will be incentivized to discriminate based on the content structure rather than only looking at the individual tokens of the text.

After surveying many NLP based text datasets, we decided to use the OpinRank review dataset (https://github.com/kavgan/OpinRank) [12], which is a dataset of cars and hotels reviews by customers. We focus on hotel reviews for our task. Our motivation for working with the reviews, in general, is that potential misuse of large scale text-generation models can be to generate fake positive or negative reviews to manipulate ratings and consumer sentiment on social media.

We generated 2000 reviews from each language model to create a total dataset of 6000 reviews. These models were given an input prompt of 300 tokens initially and 100 token long reviews were generated as we found that 100 was the average word length for reviews in the dataset.

5 Classification Problem

We map the problem of "de-anonymization" into a classification task, where our multi-class classifier will categorize the input text to be between one of the categories it was trained on.

For such a "multi-token sequence to a categorical mapping", we use LSTM neural network which is used primarily for such tasks. The final output layer will be changed according to the total number of classes in the classification task. The model architecture used is shown below:

```
-----------------------------------------------------------------
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 100, 100)          5000000
-----------------------------------------------------------------
spatial_dropout1d_1 (Spatial (None, 100, 100)          0
-----------------------------------------------------------------
lstm_1 (LSTM)                (None, 100)               80400
-----------------------------------------------------------------
dense_1 (Dense)              (None, 4)                 404
=================================================================
Total params: 5,080,804
Trainable params: 5,080,804
Non-trainable params: 0
-----------------------------------------------------------------
```

6 Experimental Results

To investigate the accuracy of the classifier in all the possible contexts, we compared each language model against each other along with classifying them against human-generated reviews as well. Finally, we train a multi-class classifier which classifies against all the models along with the human text. These are the results:

1. GPT vs GPT-2 classifier results are shown in Fig 1

2. GPT vs XLNet classifier results are shown in Fig 2

3. GPT-2 vs XLNet classifier results are shown in Fig 3

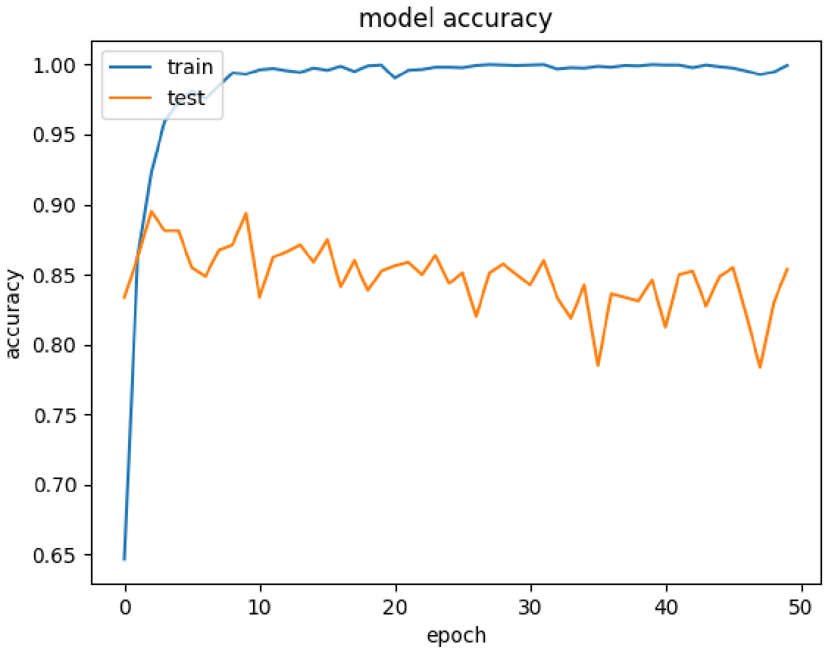
4. GPT vs GPT-2 vs XLNet vs Human-written classifier results are shown in Fig 4

7 Conclusion

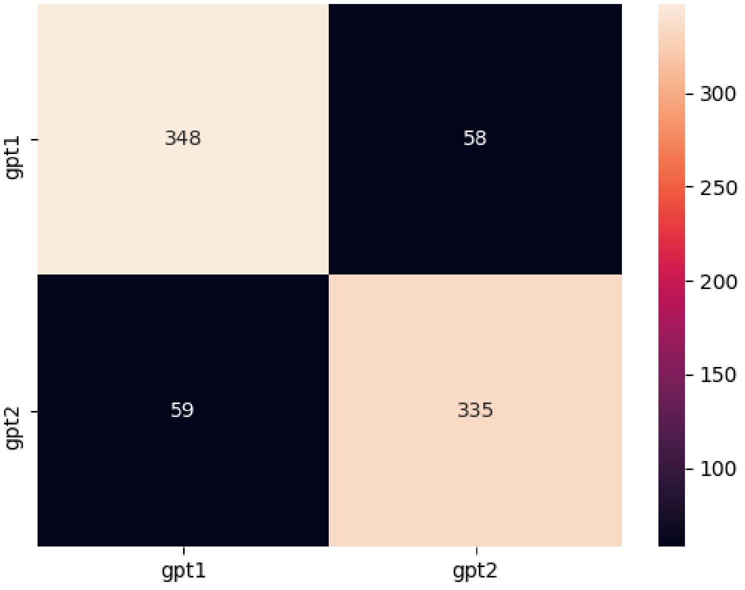
Following conclusions can be drawn from the results obtained:

1. Classifier overfits to the training data achieving accuracy close to 99.5% for almost all instances, but validation accuracy drops due to overfitting. We used dropout to add regularization but still, there seems to be a persistence of overfitting. We can explore other architectures of the classifier to alleviate this issue.

2. Higher generalization can be achieved by adding more training data and training a larger model with added drop-out layers, we decided to keep the classifier architecture relatively small owing to larger training and generation times which hinder quick prototyping and hypothesis testing.



(a) Accuracy

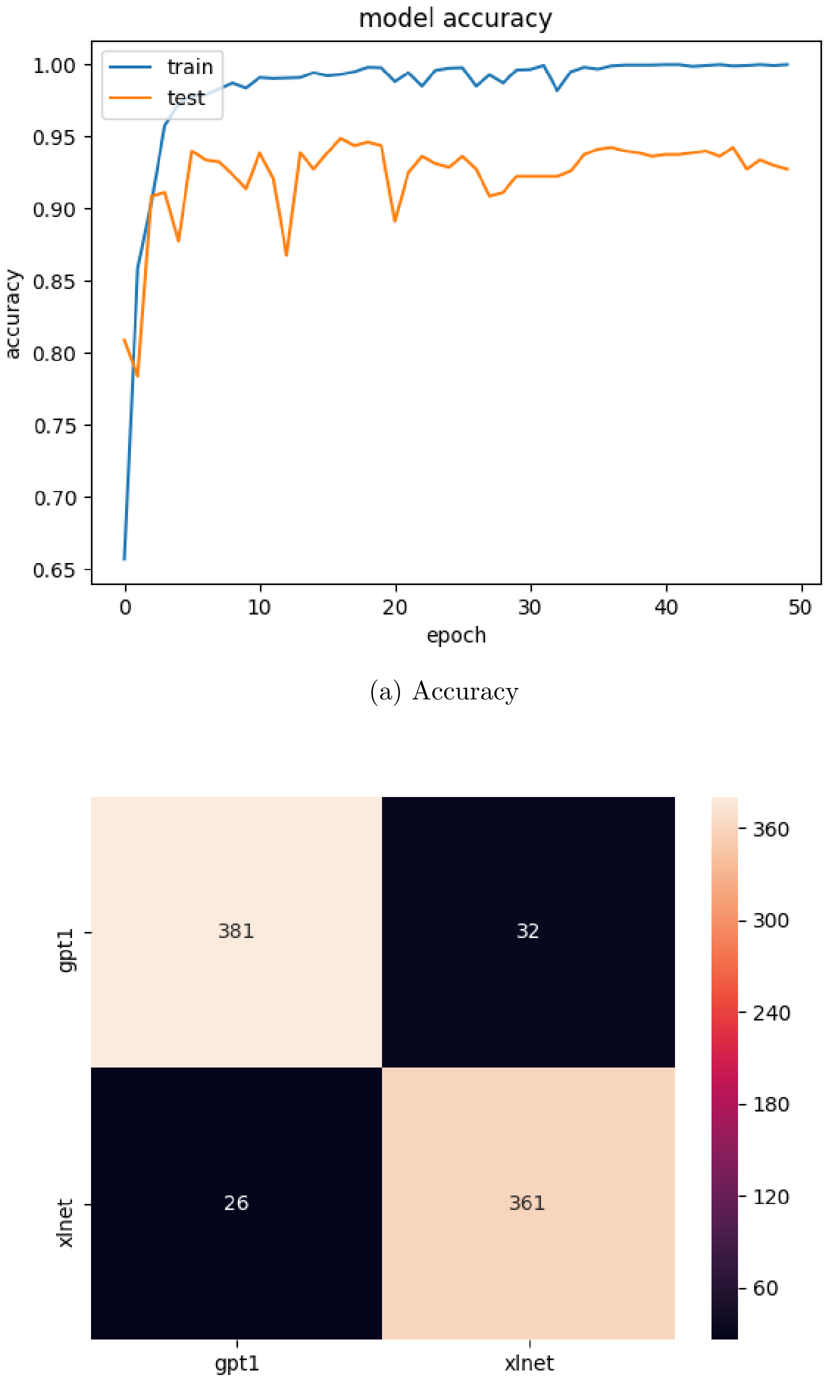


(b) Confusion Matrix

Figure 1: Classifier Performance GPT vs GPT-2. Validation accuracy is 85.12%

3. In the case of 4 class classification, the classifier performed the worst between classifying GPT - 1 and GPT - 2, which indicates the implicit text distribution is similar between models having the same architecture but different parameter scale.

4. All of the accuracies are above the simple baseline of random chance, hence it can be said classifier is able to capture the difference but as the classifier mapping function from a sequence of tokens to

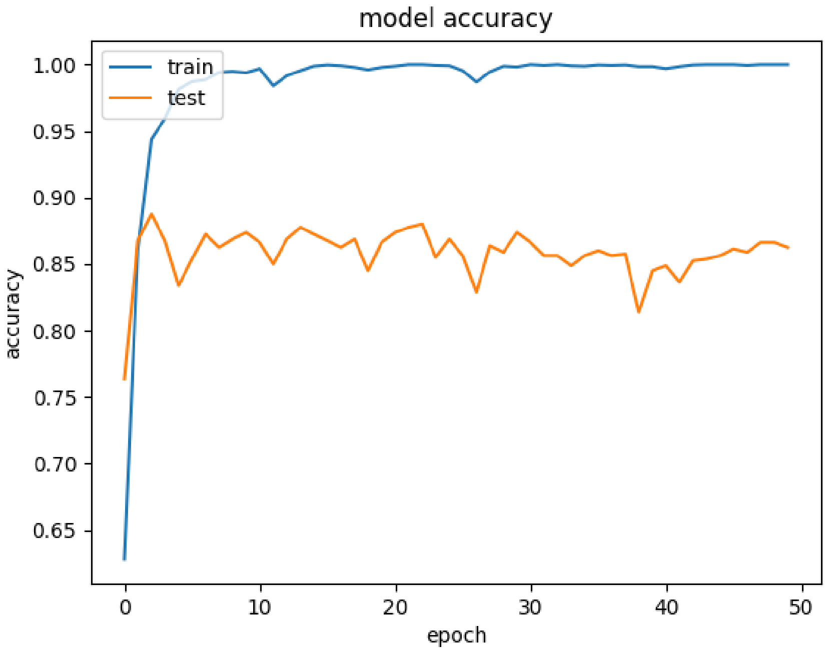


(a) Accuracy

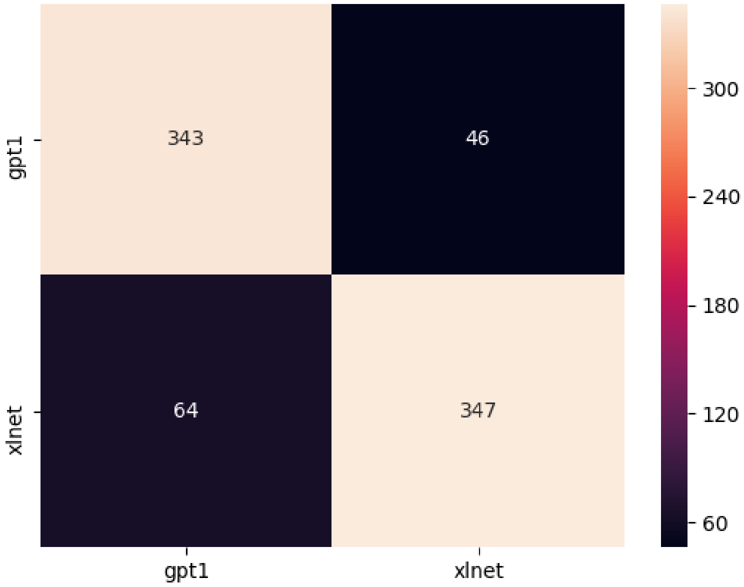
(b) Confusion Matrix

Figure 2: Classifier Performance GPT vs xlnet. Validation accuracy is 92.75%

a categorical value is highly non-linear, hence future research should be done on which will focus on the interpretability of the classification.

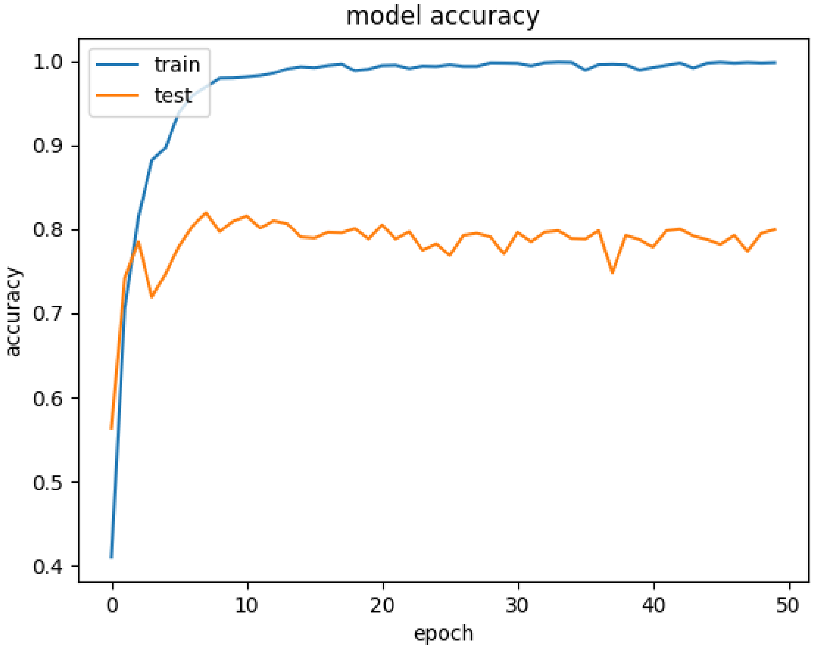


(a) Accuracy

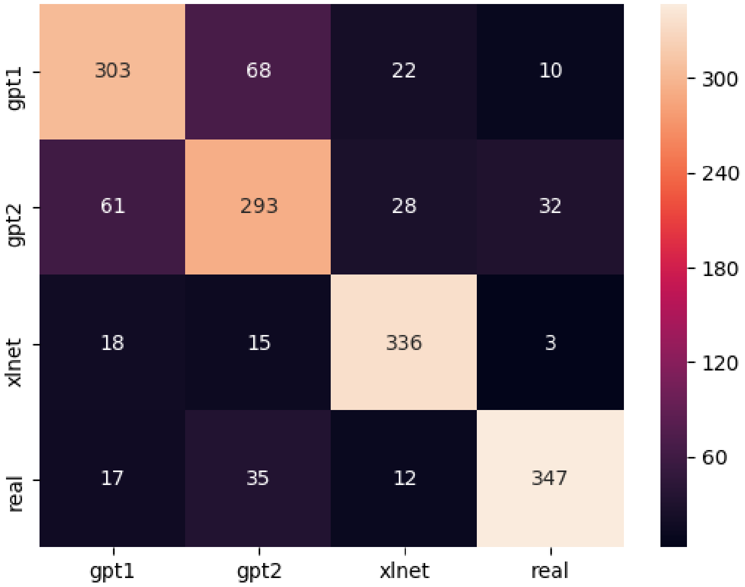


(b) Confusion Matrix

Figure 3: Classifier Performance GPT-2 vs xlnet. Validation accuracy is 86.25%



(a) Accuracy



(b) Confusion Matrix

Figure 4: Classifier Performance GPT vs GPT-2 vs xlnet vs Human-written. Validation accuracy is 80%

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, 2019.

[3] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

[4] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *arXiv preprint arXiv:1906.08237*, 2019.

[5] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," *arXiv preprint arXiv:1804.07461*, 2018.

[6] S. Afroz, A. C. Islam, A. Stolerman, R. Greenstadt, and D. McCoy, "Doppelgänger finder: Taking stylometry to the underground," in *2014 IEEE Symposium on Security and Privacy*, pp. 212–226, IEEE, 2014.

[7] C. H. Ramyaa and K. Rasheed, "Using machine learning techniques for stylometry," in *Proceedings of International Conference on Machine Learning*, 2004.

[8] A. Caliskan-Islam, R. Harang, A. Liu, A. Narayanan, C. Voss, F. Yamaguchi, and R. Greenstadt, "Deanonymizing programmers via code stylometry," in *24th {USENIX} Security Symposium ({USENIX} Security 15)*, pp. 255–270, 2015.

[9] R. Sarawgi, K. Gajulapalli, and Y. Choi, "Gender attribution: tracing stylometric evidence beyond topic and genre," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pp. 78–86, Association for Computational Linguistics, 2011.

[10] K. Surendran, O. Harilal, P. Hrudya, P. Poornachandran, and N. Suchetha, "Stylometry detection using deep learning," in *Computational Intelligence in Data Mining*, pp. 749–757, Springer, 2017.

[11] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, vol. abs/1910.03771, 2019.

[12] K. Ganesan and C. Zhai, "Opinion-based entity ranking," *Information retrieval*, vol. 15, no. 2, pp. 116–150, 2012.