CIS 6930: Privacy & Machine Learning

Mid-Semester Project Report: Deep learning-based approaches for stylometric De-Anonymization for model authorship

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

Kunwardeep Singh
kunwardeep.singh@ufl.edu

November 14, 2019

1 Progress & Preliminary Results

Phase 1: Training data-set generation:

1. We picked the top 4 state of the art language models in the literature, 1. OpenAI GPT 2. GPT2 3. XL-NET 4. Salesforce CTRL. After looking and testing different open-source implementation of these models, we decided to use huggingface transformer library [1] 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.

2. 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.

3. After surveying many NLP based text datasets, we decided to use the OpinRank review dataset [2], 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.

Phase 2: Classification task

1. For the initial prototype, we decided to perform a binary classification task between two models with similar architecture but different parameters, GPT -1 and GPT. We created a sample dataset of 100 model generated reviews given same initial input of hotel reviews in the city of Chicago.

2. We created a lstm neural-network with the following architecture to train a neural network to differentiate between two models. We achieved a test accuracy of 68% and a training accuracy of 91% for the initial training iteration. This indicated a high overfitting of the classifier. We concluded we needed more training data to sufficient comment on the performance as well as the generalization capacity of our lstm model.

```
Layer (type)                 Output Shape              Param #
=================================================================
embedding_4 (Embedding)      (None, 50, 32)            160000
_______________________
lstm_4 (LSTM)                (None, 100)               53200
_______________________
dense_4 (Dense)              (None, 1)                 101
=================================================================
Total params: 213,301
Trainable params: 213,301
Non-trainable params: 0
```

3. Total training and text generation at large scale take a very long time on local machines, so we are using HiperGator for large scale text generation and training of classifiers, which should complete within a week.

2 Future Work

1. Currently, we are looking at different classification neural networks other than lstm, which can give better performance for this particular task.

2. After creating sufficiently large datasets, we will train our classifier to obtain performance metrics like accuracy, precision, recall, etc. and check for overfitting. We are also investigating the relationship between the architecture of the text generation model and the de-anonymization success rate for that particular model.

3. We are planning to do pairwise binary classifiers and all in one k class classifier to check is there any significant performance difference between specific pair of models.

4. We are also planning to add human-generated text as an additional category and check how differentiating it is compared to other model-authored texts.

References

[1] 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.

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