CIS 6930: Privacy & Machine Learning

Project Proposal: 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

September 27, 2019

1 Introduction

With rapid advancement in Natural Language Processing and Deep learning, the end-to-end pipeline for many natural language-related tasks is possible. With new architectures for Language modeling achieving very impressive results in Natural Language Generation, neural networks can output coherent text given some textual prompt. 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 if they are deployed with a malicious intent.

In our project, we are proposing to develop a Deep-learning based classifier that can classify text generated by the top 4 State of the Art "Language Modelling" open-source models. Our goal is to research how much effect the architecture of a neural network has on its style, if any. We will add another class of human-generated text and examine how much stylistic difference is present between current SoTA Language Models and human-written text. Also, we are planning to investigate the effectiveness of current Stylometry methods on Neural Network generated text to find out if current techniques still work effectively or we need some modified approach for tackling this problem.

2 Background and 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. [1], Ramyaa et al. [2]) But in these methods, the features based on which classification is done are hand-crafted and extracted manually. Also, Caliskan-Islam et al [3] 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. [4], Surendran et al.[5]).

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.

3 Proposed Approach & Plan

We plan to approach our problem in the following phases:

1. Model selection
   (a) There are many open-source Natural Language Generation models trained on a specific dataset and specific task with different architecture. We are planning to focus on Language Modelling neural networks that are tasked to predict the next word in a sentence given a sequence of prior occurring word tokens (eg. GPT-2). We plan to pick Top 4 (this top-k parameter can vary) State of The Art models in Language modeling with papers and code can be seen here for reference - https://paperswithcode.com/task/language-modelling. Also, open-source code for various these models is available here: https://github.com/huggingface/transformers. We examine and select the best performing models and also consider architecture and other factors while finalizing them as classes in our classification task.
2. Training data gathering
   (a) OpenAI has open-sourced its GPT-2 output dataset (https://github.com/openai/gpt-2-output-dataset). Similarly, we are planning to run the pre-trained models to generate prompts and collect sufficient data and label them to assemble training data required to train our classifier.
   (b) Human-generated text datasets are readily available. (for eg. https://github.com/niderhoff/nlp-datasets)
3. Design and development of neural network-based classifier.
   (a) During and after the training data gathering phase we are planning to start the development of our classifier neural network.
   (b) We are planning to build an lstm based classifier, but we will finalize architecture and hyper-parameters after some initial experiments and preliminary findings.
4. Evaluation and findings
   (a) After sufficient training, We will quantitatively examine the accuracy of our classifier and determine the extent to which stylometric difference is present in different neural networks.
   (b) Also, we will qualitatively examine the extent to which architecture, content of text (sports, politics, science, art) determines stylometric differences between various models.

4 Timeline

Milestone	Target Date
Evaluate and select models based on performance	10/11/2019
Gather and clean the dataset	10/18/2019
Start developing the classifier	10/25/2019
Submit mid semester project report	10/31/2019
Incorporate feedback received in the project	11/08/2019
Optimize the classifier after several iterations	11/22/2019
Start evaluation and gather results	11/29/2019
Final Project Report	12/11/2019

References

[1] 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.

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

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

[4] 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.

[5] 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.