Milestone 2: Transformer-based vs Ensemble Architecture for PII Identification

Innocent Farai Chikwanda
innocent.chikwanda@ashesi.edu.gh

Tendai Terrence Machaya
tendai.machaya@ashesi.edu.gh

Introduction

The growing digitalization of businesses, online learning, and governments has led to an exponential increase in online data collection, a significant portion of which comprises Personal Identity Information (PII)[2]. PII encompasses sensitive information, like individual names and social security numbers, and non-sensitive identifier information, such as nationality. The mishandling of PII, whether through inadvertent disclosure or malicious intent, poses severe risks to individuals' privacy and can lead to financial or reputational losses for organizations [1].

Given these challenges and the increasing importance of data security and privacy, there is a pressing need for effective PII detection and protection mechanisms. Natural Language Processing (NLP) techniques have emerged as valuable tools in this domain, enabling organizations to identify and mask PII efficiently. However, existing PII detection tools, while effective to some extent, often need help with limitations such as false positives and the inability to handle unstructured data effectively.

Proposal

We have chosen to implement, test and compare three main types of natural language processing models that have been proven highly effective in solving Named Entity Recognition(NER) problems in natural language processing. These best performing approaches include an enhanced bidirectional LSTM, Transformer-based architecture, and an ensemble architecture that combines NER models and rule-based approaches [2]. These three approaches are particularly suited for this problem due to their ability to capture the long-term contextual relationship between tokens which is very important in identifying easily confusable PII like names. We are aware that these models have different strengths and we are curious to test out these three approaches on our dataset selecting the best performing model across several metrics, particularly f5-score, recall, precision and accuracy. While similar work has been done in the past with ensemble model proving better than the transformer-based model BERT[2] on all metrics, we hope to utilize a comparatively more capable but finetuned GPT-4 model[4] using the Kaggle dataset augmented with some externally sourced PII datasets, preprocessed to the same format as the OpenAI fine tuning dataset specifications. We also hope to take full

advantage of the rule-based classification offered by our ensemble model, SpaCY, in addition to its NER model classification for unambiguous PII like phone numbers. Implementing these models would suffice to produce a high-performing model either using an exclusive model or ensemble of models. We will most likely make use of the Weights and Biases platform to track all the modell experiments on this task in a more effective way due to their robust MLOps Pipeline.

References

1. IBM Security. "Cost of a Data Breach Report 2021." Accessed from https://www.ibm.com/security/digital-assets/cost-data-breach-report/#/.
2. Lin, T. J., & Abhishek, N. V. (Year). Personal Identity Information Detection using Synthetic Dataset. Information and Communication Technology Cluster, Singapore Institute of Technology.
3. Amin, S., Neumann, G. (Year). T2NER: Transformers based Transfer Learning Framework for Named Entity Recognition. Department of Language Science and Technology, Saarland University, Saarbrucken; Multilinguality and Language Technology Lab, DFKI GmbH, Saarbrucken. {saadullah.amin, guenter.neumann}@dfki.de
4. Grishina, A. (2024, January 5). GPT-3 vs. BERT: Ending The Controversy. SoftTeco. https://softteco.com/blog/bert-vs-chatgpt
5. Instrumenting Weights and Biases: PII data detection https://youtu.be/w4ZDwiSXMK0?si=tMhLlMuugGec8TjI