Project Plan for Engineering Degree Projects
Department of Computer Science

General Information

Title:	Detection and Replacement of Sensitive Information in Free-Text
External company:	TietoEvry

Persons involved

Student	Felix Böhlin	fb222ur@student.lnu.se

Supervisor:	Ola Flygt
External supervisor:	Mattias Holmgren

Background

Individuals are becoming more concerned about their digital footprint, trying to protect personal and sensitive information from unauthorized access. A response to businesses collecting more personal data for unintended purposes without online users approval[1]. Organizations prioritize data protection to build a reputation of reliability and integrity for its customers[2]. They also do it because they have to, under the regulations such as the General Data Protection Regulation (GDPR) which set the standard for secure handling of personal information. If not handled correctly could hefty fines and legal repercussions be the outcome. Manually going though data and erasing sensitive information is not practical and resource heavy. Therefore, is this project a collaboration with TietoEvry which in their case is collaborating with Friends to come up with a AI algorithm that can anonymize any sensitive and personal information in unstructured text. TietoEvry is a company at the forefront of technology and innovation that creates digital solutions for businesses and communities.

Related Work

"Using Machine Learning to Help Detect Sensitive Information"[3] is about detecting and flagging any document that includes sensitive information. This project also aims to detect sensitive information and does so by filtering out any non-informative words from the texts. Then uses the random forest classifier model to be able and single out the documents that are sensitive. However, the document it self doesn't have to be sensitive but there could be segments within the document that includes personally identifiable information (PII). To detects this data do they use regexs which can find patters in the text. In contrast, this project at TietoEvry aims to advance the detection of sensitive information in unstructured text not just at the document level but also within specific segments. The goal is then to replace the PII with dummy data and not just flag the whole document.

"Automated detection of unstructured context-dependent sensitive information using deep learning"[4] is about detecting sensitive information in unstructured data using deep learning techniques. These techniques are Word Embedding for identifying context in the data and Neural Networks for the classification. They also compared the accuracy to other models like SVM and Bayers with their input data coming from tweets and images. The deep learning methods showcased significantly better results compared to the traditional machine learning models. This advantage, however, comes with a trade-off where deep learning methods are more resource heavy. Despite this, the potential benefits make deep learning an appealing choice for this project. To avoid these challenges of developing new models will this project explore existing large language models (LLMs) with tuning to match the problem in hand.

Problem formulation

The problem lies in efficiently identifying and anonymizing sensitive information within unstructured text. This challenge is increased by the requirements of regulations such as the GDPR, which demands a high degree of accuracy and reliability in handling personal data. Current non-AI methods for manual or semi-automated solutions are not scalable, prone to errors, and resource heavy. This project, in collaboration with TietoEvry aims to

minimize these problems with machine learning techniques to automate the process of detecting and anonymizing sensitive information.

This project could see limitations in computational resources since training machine learning models to handle complex tasks takes large volumes of data especially with large volumes of unstructured text. TietoEvry will provide the datasets which removes the time and complexity to gather this information which would be challenging since collection of sensitive information is not easy.

1. Develop a agnostic platform solution that can be runnable on a local setup and cloud platforms.

2. Develop a reliable, cost-efficient solution for detecting and replacing sensitive information in text.

Motivation

Developing and refining machine learning models for detecting and anonymizing sensitive information in unstructured text are crucial. It will contribute to the advancement for natural language processing (NLP) which is the method for building models that can manipulate human language or data that resembles human language [5]. In today's digital era, where data is increasingly at the core of our daily lives, the risks associated with the misuse of personal information, such as identity theft, financial fraud, and privacy violations is escalating.

Automated tools for the detection and anonymization of sensitive information are essential for individuals to retain control over their personal data and ensure their privacy. From an industry perspective, organizations across various sectors are gathering large amounts of data at an fast pace. This unstructured data, great for insights that can better innovation, and secure a competitive edge. However, it also presents significant challenges if the sensitive information it contains is not correctly protected.

The repercussions of data breaches extend beyond direct financial losses to include long-term reputational damage that can harm customer trust. Implementing automated and efficient mechanisms to identify and anonymize

sensitive information is important for ensuring data protection regulations, like GDPR. This would not only avoid financial penalties but also strengthen brand integrity by showcasing a strong commitment to customer's privacy.

Milestones and Time Plan

M1	Have a finished road map, knowing which models and practices to use	April 10
M2	Turn in half-way report	April 24
M3	Have the first working model	May 5
M4	Test the final system and refine it to strike the right balance between sensitivity and specificity	May 15
M5	Have a finished agnostic platform solution	May 20
M6	Oral presentation and opposition	May 30

Method

For the project will a combination of Literature Review and Controlled Experiments be used to investigate, develop, test and deploy the solution. The methods are chosen to align with the milestones ensuring the understanding and completion of the problem.

Literature Review

Literature review will be conducted to explore the selection of deep learning and machine learning models. This method aims to understand existing approaches to the problem, making the selection of the most efficient and cost-effective models. It will focus on identifying best practices and understanding current challenges to address them effectively. This method is directly related to Milestone 1, and will also help the development of the machine learning models.

Controlled Experiments

Controlled Experiments will be used to test the models ability to detect and replace sensitive information from text. Basically evaluating the performance

of the model. This method is important to be able and reach Milestone 4. The metrics that will be used for the testing are precision, recall and F1 score. Precision will measure a models positive predictions, meaning when a model accurately identifies a sensitive text bit. Recall will measure all positive outcomes, meaning the models ability to identify sensitive text bits compared to all sensitive instances in the text. The F1 score will provide a value that shows how good the balance between precision and recall is, measuring the models overall accuracy.

References

[1] Christine Prince. Do consumers want to control their personal data? empirical evidence. *International Journal of Human-Computer Studies*, 110:21–32, 2018.

[2] Donal Tobin. What is data privacy—and why is it important?, 2024. [Online], https://www.integrate.io/blog/what-is-data-privacy-why-is-it-important/.

[3] Renae Kang. Using machine learning to help detect sensitive information, 2023. [Online], https://blog.developer.adobe.com/using-machine-learning-to-help-detect-sensitive-information-5bfb32eeb34e.

[4] Hadeer Ahmed, Issa Traore, Sherif Saad, and Mohammad Mamun. Automated detection of unstructured context-dependent sensitive information using deep learning. *Internet of Things*, 16:100444, 2021.

[5] n.d. A complete guide to natural language processing, 2023. [Online], https://www.deeplearning.ai/resources/natural-language-processing/.