Research Paper E-ISSN NO : 2455-295X | VOLUME : 9 | ISSUE : 6 | JUNE 2023

DETECTION OF PRIVATE DATA IN MOODLE'S TEXT-BASED FIELDS

BEATA GANCEVSKA [1] | PAULIUS NOMGAUDAS [2]

[1] DEPARTMENT OF INFORMATION SYSTEMS, VILNIUS GEDIMINAS TECHNICAL UNIVERSITY, VILNIUS LT-10223, LITHUANIA.

[2] DEPARTMENT OF INFORMATION SYSTEMS, VILNIUS GEDIMINAS TECHNICAL UNIVERSITY, VILNIUS LT-10223, LITHUANIA.

ABSTRACT:

Moodle is the most popular and widely used learning management system, but it collects and stores a lot of personal information. The use of such data for analysis, development or system testing raises significant privacy concerns. This paper investigates a combination of approaches to anonymizing specific attributes in the *Moodle* database. The newly proposed method uses named entity recognition techniques to find personally identifiable information, while more traditional methods can be used for other private data (IP, e-mail, etc). The obtained results identify that Lithuanian language-specific data is lacking accuracy in the named entity recognition area (recall for names and company titles reaches up to 94%, however for different address elements recall values reach just up to 81%), while common format data is easily recognized even without the usage of machine learning solutions.

KEYWORDS:

DATA ANONYMIZATION, NATURAL LANGUAGE PROCESSING, NAMED ENTITY RECOGNITION, *MOODLE*.

INTRODUCTION

Moodle learning management system (LMS) is widely used but does not have a big variety of data analytics functions and plug-ins. Sometimes Moodle data is shared with another system to assure data analysis possibilities. This leads to a possible data privacy problem as third parties might get sensitive student data. While the traditional student profile data can be easily removed or updated to random data, the text field data, where students and teaching staff communicate, might also include some personal data. There are no easy solutions for identifying privacy-sensitive data in *Moodle* data.

This paper aims to investigate the approach to anonymizing text-based data in *Moodle* using NER methods and pattern-based solutions. The proposed method identifies and anonymizes sensitive text data attributes in the *Moodle* database while preserving attribute format, symbol length and usability. Anonymous data can be used for testing and development purposes.

NER is a technique used to find and extract named entities. This technique identifies entities in the text such as people's names, surnames, locations, organizations, time, and other information (Jing, Aixin, Jianglei & Chenliang, 2022). Qiu et al. (2019) claim that NER methods can be classified into Rule-based NER, Machine Learning-based NER, and Hybrid NER.

Rule-based NER is the first method applied to named entities (Eftimov, 2017). This technique involves manually creating rules that will be used to identify named entities in text. This method also uses regexes (Perera et al., 2020). With regular expressions, it is possible to describe a pattern where the required data will be found.

Another widely used technique in the NER field is artificial intelligence. In this case, finding and categorizing identifiers is done using machine learning algorithms. Machine learning methods such as *Hidden Markov Models* (HMM), *Support Vector Machines* (SVM), and *Conditional Random Fields* (CRF) are employed for the recognition of named entities (Jing, Aixin, Jianglei & Chenliang, 2022).

The third category is hybrid NER. Hybrid NER methods combine rule-based and machine learning-based approaches to improve named entity recognition accuracy. Hybrid systems are more accurate than individual systems (Goyal, Gupta & Kumar, 2018).

Solutions for data anonymization in *Moodle* exist. The *Make Anonymous*, *Moodle User Anonymizer* plugins focus on specific database fields, but not private data detection in text fields. *Anonymise* plugin anonymizes a variety of data types, such as activities, course categories, courses, user data, and other sensitive information, across an entire *Moodle* site. However, there is currently no solution available for recognizing and anonymizing identifiers specifically in the Lithuanian language. Language specifics play a significant role and must be considered (Biesner et al. 2020). It is, therefore, necessary to research different methods of anonymizing *Moodle* data to detect Lithuanian user's personal information.

1. MATERIALS AND METHODS

1.1. PROPOSED PRIVATE DATA RECOGNITION AND ANONYMIZATION METHOD

The proposed private data (PD) recognition and anonymization method consists of two main steps (see Figure 1). The first step involves the recognition of specific data attributes in the text that might contain private data. This includes data such as personal codes, telephone numbers, IP addresses, dates, email addresses, and postal codes, which can be identified using regexes. NLP searches for first names and surnames, companies, and place names.

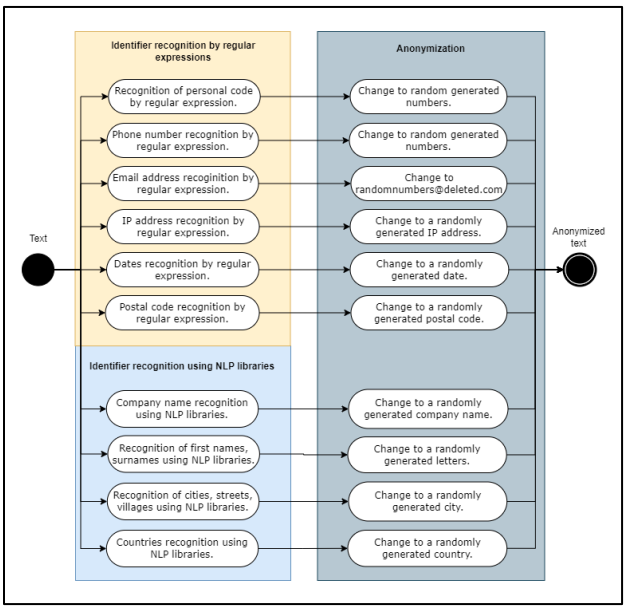

FIGURE 1. PROPOSED PD RECOGNITION AND ANONYMIZATION METHOD.

For easier experimentation and application of the proposed solution, the *Moodle* plugin was created. It has two main components: data recognition and anonymization. The plug-in settings allow the selection of *Moodle* data fields on which the methods should be applied and automate data anonymization.

1.2. DATASET CREATION FOR DATA ANONYMIZATION TESTING

The proposed data anonymization method is tested by anonymizing *Moodle Chat* activity. To label data accurately, templates were generated and used for *Moodle Chat* activity record generation. The chat messages were generated in Lithuanian.

The created dataset contains 1000 messages where multiple private data elements (unique personal identification numbers used in Lithuania, email addresses, phone numbers, dates of birth, first and last names, IP addresses, postal codes, company names, and residential addresses) are presented in each of them. Figure 2 shows the quantities of 3021 personal data elements for each category among 1000 messages.

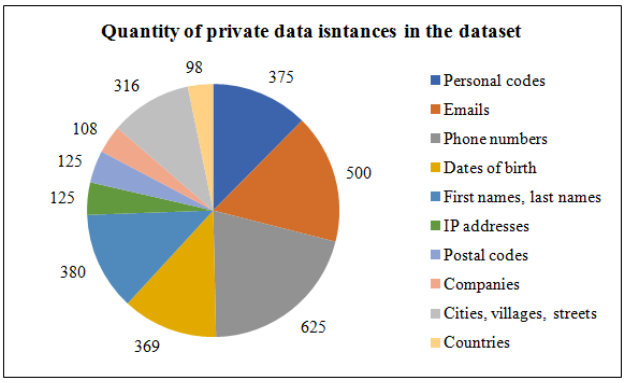

FIGURE 2. DISTRIBUTION OF PD ELEMENTS IN THE DATASET

1.3. PERSONAL DATA DETECTION EXPERIMENT SETUPS

As presented in Figure 1, the experiments with PD detection were organized into two parts: application of identifiers using regular expressions (Rule-based NER) and experiments with NER application (machine learning methods). By combining both parts, a hybrid NER was implemented.

In each of those parts, the whole dataset was used for testing, not training. This is explained by the fact that rule-based NER did not require training, just rule definition. While for machine learning experiments, existing NER models were applied (*Spacy* and *Polygot*). The models have Lithuanian language support and can be used for Lithuanian language NER.

In data anonymization, the most significant metric is *Recall*, which identifies the proportion of correctly recognized elements among all private data instances. However, in the experiment, *Precision* and *F-score* were also calculated. Those metrics allow to analyze the wider view, including false positives as well.

2. RESULTS OF PRIVATE DATA DETECTION EXPERIMENTS

2.1. IDENTIFIERS RECOGNITION USING REGULAR EXPRESSIONS

Using regular expressions, personal codes, email addresses, phone numbers, dates of birth, IP addresses, and postal codes are recognized. The experiment included iterative modification of regular expressions while the final regular expression for each data field was obtained. A 100% accuracy result was achieved (*Precision*, *Recall*, and *F-score* values are also 100%). This indicates this approach is suitable for Lithuanian systems and patterns and produces no false positive or false negative cases. However, defining regular expressions for textual data fields led to failure as accuracy metrics were extremely low. Therefore machine learning NER was used for the rest of the private data elements.

2.2. IDENTIFIERS RECOGNITION USING NLP SOLUTIONS

The *Spacy* and *Polyglot* libraries were used for the recognition of first names, surnames, company names, countries, cities, streets, and villages were performed. Those libraries support the Lithuanian language, by using appropriate language models. As the most important is to find all private data elements in the text, the *Recall* scores are presented in Figure 3. It illustrates that none of the models were capable to identify all the PD instances in the text. As well it shows very different *Recall* scores for different PD elements: *Spacy* performed not badly in all categories, while *Polygot* combined the country and address data into one category, but demonstrated worse results in all categories, especially first names, surnames and companies identification below 20% among all actual entities.

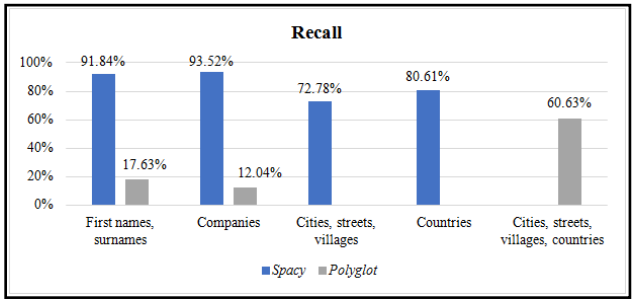

FIGURE 3. RECALL IN IDENTIFYING PD ELEMENTS.

Similar results were demonstrated for *Precision* as well. *Spacy* library has the highest *Precision* in identifying PD elements (from 86% to 91% for separate categories). Using the *Polyglot* library, the highest *Precision* was achieved in first and last name recognition (up to 86%), but company name recognition was achieved with only 35% *Precision*.

The average *F-score* for all categories of PD entities for the *Spacy* library is 87%, while for the Polyglot library, it reaches just 47%. It is shown here that the *Spacy* library is more suitable for Lithuanian NER.

3. DISCUSSION

This study showed the effectiveness of different methods for identifying and anonymizing PD in *Moodle* chat messages. Regular expressions were fully justified for certain data elements, while NLP libraries lack accuracy in recognizing identifiers. For all PD elements, the hybrid NER application (regular expressions + *Spacy* library) provided a 96% *F-score*. The *Recall* metric is more important in data anonymization as it indicates the ratio between correctly identified PD elements and the total number of existing PD elements in the dataset. The *Recall* score for the hybrid solutions reaches 95%. This illustrates a very similar result for all metrics.

The research concludes that the use of regular expressions is highly effective at identifying identifiers, such as personal codes, email addresses, phone numbers, dates of birth, IP addresses, and postal codes. Therefore, regular expressions are a reliable method of identifying these data attributes.

The use of NLP libraries yielded mixed results. The *Spacy* library has high *Precision* and *Recall* in identifying PD elements, such as people's first and last names. However, it showed moderate recall for cities, streets, and villages. The *Polyglot* library showed moderate *Recall* in recognizing place names, but its *Precision* rate was low. The *Polyglot* library also had a low precision rate for company name recognition, 35.13% precision.

The *Spacy* library had the highest *F-score* for first and last name recognition, over 90%. Therefore, *Spacy* is a suitable method for identifying PD elements in *Moodle* chat messages. However, improvement is needed to increase *Recall* for other types of identifiers.

Overall, this study demonstrated that the NER solutions are not 100% accurate for the Lithuanian language in *Moodle* chat messages. This is because Lithuanian language is more complex, having a variety of forms of words, etc. Therefore, further research is needed to improve the recognition and anonymization of specific PD attributes for the Lithuanian language.

REFERENCES

1. Biesner, D., Ramamurthy, R., Lübbering, M., Fürst, B., Ismail, H., Hillebrand, L., ... & Sifa, R. (2020). Leveraging Contextual Text Representations for Anonymizing German Financial Documents. Proc. Knowledge Discovery from Unstructured Data in Financial Services. AAAI.

2. Eftimov, T., Koroušić Seljak, B., & Korošec, P. (2017). A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. PloS one, 12(6), e0179488.

3. Goyal, A., Kumar, M., & Gupta, V. (2017). Named entity recognition: applications, approaches and challenges. International Journal of Advance Research in Science and Engineering, 35(5), 482-489.

4. Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, 34(1), 50-70.

5. Perera, N., Dehmer, M., & Emmert-Streib, F. (2020). Named entity recognition and relation detection for biomedical information extraction. Frontiers in cell and developmental biology, 673.

6. Qiu, Q., Xie, Z., Wu, L., & Tao, L. (2019). GNER: A generative model for geological named entity recognition without labeled data using deep learning. Earth and Space science, 6(6), 931-946.