PReDUS: A Privacy Requirements Detector From User Stories

Francesco Casillo[1], Vincenzo Deufemia[1] and Carmine Gravino[1]

[1]*Department of Computer Science, University of Salerno, Via Giovanni Paolo II, 132, Fisciano(SA),84084, Italy*

Abstract

In the context of requirements engineering, stakeholders are often unaware of identifying and managing privacy and security requirements. The purpose of this paper is to present a tool, namely PReDUS, for the detection of privacy content from user stories. The core of the tool is the use of deep learning algorithms that exploit Natural Language Processing techniques and linguistic resources.

Keywords

User Stories, Natural Language Processing, Deep Learning, Transfer Learning

1. Introduction

Identifying non-functional requirements (NFRs) from stakeholders during requirements engineering (RE) phase can be a problematic activity due to several factors [1]. Failure to take care in documenting and defining these requirements can lead to defects in software development [2]. Privacy is one of the NFRs that has become more important in recent years, in part because the needs of businesses increasingly require the protection and safeguarding of their data [3]. Although privacy requirements are inherent in the software development process, stakeholders are often unable to recognize them from customer requirements [4].

In order to provide an automated solution for the detection of privacy content from requirements defined as User Stories, we introduce PReDUS, a tool exploiting recent Natural Language Processing (NLP) technologies to extract features which are then used by a convolutional neural networks (CNNs) based model, obtained by employing Transfer Learning technique [5].

2. PReDUS's approach

PReDUS is a web application that aims to provide some insights about the privacy requirements contained in a User Story (US) [6]. Figure 1 shows the main components of PReDUS to identify privacy content. The text of the input US is first processed by the NLP Toolkit in order to capture both the grammatical structure of the text and the meaning being conveyed. Then, the output of this phase becomes input to a Transfer Learning model, which allows to involve

NLP4RE'22: 5th Workshop on Natural Language Processing for Requirements Engineering REFSQ'22 Workshop, Aston, Birmingham, UK, March 21-24, 2022

fcasillo@unisa.it (F. Casillo); deufemia@unisa.it (V. Deufemia); gravino@unisa.it (C. Gravino)
0000-0003-4869-8068 (F. Casillo); 0000-0002-6711-3590 (V. Deufemia); 0000-0002-4394-9035 (C. Gravino)

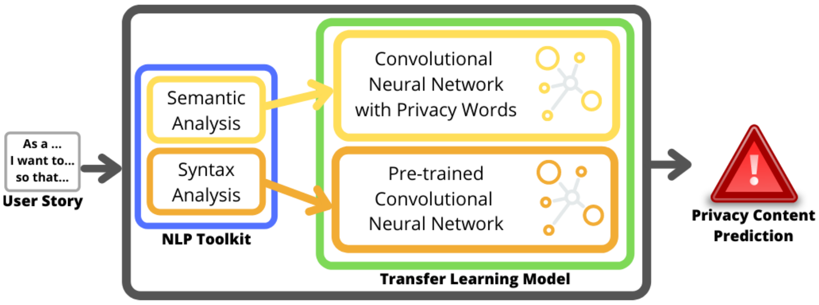
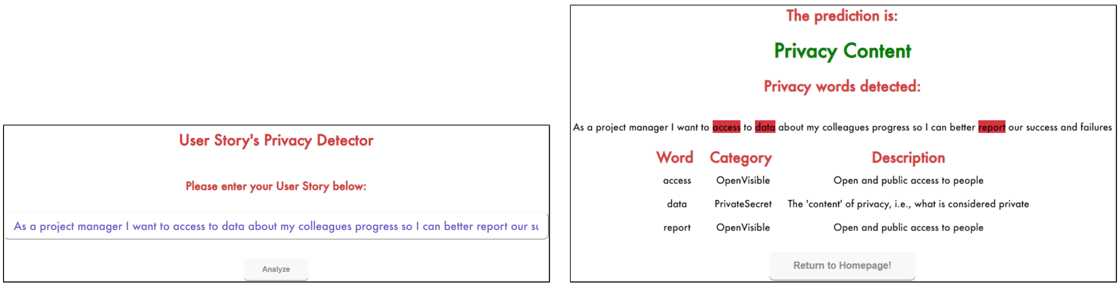

As a ...
I want to...
so that...
User Story
Semantic Analysis
Syntax Analysis
NLP Toolkit
Convolutional Neural Network with Privacy Words
Pre-trained Convolutional Neural Network
Transfer Learning Model
Privacy Content Prediction

Figure 1: PReDUS prediction process.



User Story's Privacy Detector

Please enter your User Story below:

As a project manager I want to access to data about my colleagues progress so I can better report our su

The prediction is:

Privacy Content

Privacy words detected:

As a project manager I want to access to data about my colleagues progress so I can better report our success and failures

Word	Category	Description
access	OpenVisible	Open and public access to people
data	PrivateSecret	The 'content' of privacy, i.e., what is considered private
report	OpenVisible	Open and public access to people

(a) Input data view.

(b) Results view.

Figure 2: PReDUS User Interface

two CNNs to accomplish a semantic and syntactic analyses, respectively. In this way, both syntax and meaning are analyzed to detect privacy content. In spite of a simple user interface (see Figure 2), the tool is highly recommended for stakeholders with a few experience in the identification of privacy content: the detection of this NFR in the requirements definition phase could anticipate significant issues that can lead to software malfunctions [7].

I/O Data. The algorithm designed to make PReDUS work requires a US as input, whose format is, as designed in literature [8], as the following:

“`As a` [*role*]`, I want to` [*feature*]`, so that` [*reason*]”.

As an example: “`As a project manager I want to access to data about my colleagues progress so I can better report our success and failures.`”

The US to be analyzed will be texted into the form provided in the first page of the web application (see Figure 2a). The output of PReDUS is the prediction of the deep learning algorithm developed to detect privacy content, supported by further useful information like the privacy words identified, the categories they belong to, and the description of the category to specify why the considered word is related to privacy matters (see Figure 2b).

Preprocessing. SpaCy[1] is used to extract features exploited by the algorithm to make the prediction. In particular, the algorithm uses two types of analysis, summarized as follows:

1. **Syntax Analysis.** The input US is first tokenized by using `nlp` object[2] of spaCy. It is essentially a pipeline of several text pre-processing operations applied to extract:
   - **Entities**[3]. SpaCy provides an efficient statistical system that can assign labels to individual tokens or groups of tokens that are contiguous. It can recognize a wide range of named or numeric entities, which include people, organizations, languages, events and so on.
   - **Parts of Speech**[4]. The Part of speech (POS) tagging is the process of marking a word in the text to a particular part of speech based on both its context and definition. In brief, it is the process of identifying a word as nouns, pronouns, verbs, and so on.
   - **Dependencies**[4]. Dependency parsing is the process of analyzing the grammatical structure of a sentence based on the dependencies between its words. Words are replaced by tags, called dependency tags, that represent the relationship between two or more words.
2. **Semantic analysis.** It is carried out on the individual terms of the US to search for terms strongly related to privacy issues aiming to reinforce what has been done in the previous step and to expand the number of features which will be used for privacy content predictions. The dictionary proposed in [9], and specifically developed for privacy content analysis, has been used to facilitate the search of privacy-related terms, and to obtain the privacy category each term belongs to.

Privacy Detection. The CNNs based model to predict whether the US contains or not privacy content is built using Transfer Learning, an advanced deep learning technique that consists in reusing the knowledge developed by an algorithm to solve a task and applying it to a different but related problem (see [4] for details). In particular, a neural network developed for the detection of privacy disclosures in an unstructured text is used to increase the number of features exploited by our prediction model. Further details about the implemented CNNs can be found in [4]. The classification of USs is based on the assumption that if the US contains privacy disclosures and contains privacy-related words, then the US is highly related to privacy issues. The result tell us if the considered US contains or not privacy content.

3. Demo plan

Environment Configuration. To use PReDUS you must start the server by running a Python[5] script. To make the script work, you need to install nine libraries that are useful for both the User Interface and the US analysis. In particular, Flask and IPython were used to create the UI, while the remaining libraries allowed us to handle data (Pandas), to process them (NLTK,

[1] https://spacy.io

[2] https://spacy.io/usage/models

[3] https://spacy.io/usage/spacy-101#annotations-ner

[4] https://spacy.io/usage/linguistic-features

[5] https://www.python.org/

spaCy, Numpy), and to get the model work (Tensorflow, Pickleshare, Keras). PReDUS and the libraries are available on Github[6].

Privacy detection execution. As shown in Figure 2a, first the user is asked to enter a US in the text box and then he/she can start the privacy content detection process by pressing the "Analyze" button. The application server processes the US given as input, performs semantic and syntactic analyses whose results are used as input to the CNNs based Transfer Learning model, whose output is the prediction regarding privacy content. In particular, as shown in Figure 2b the user is informed about the presence of privacy content, the identified privacy terms, the categories those terms belong to and its description to explain why that terms are related to privacy matters. Further explanations about privacy categories are explained in [9].

Usage examples. To demonstrate the usefulness and effectiveness of PReDUS we show three use case scenarios from the Web application domain. The first US given in input to PReDUS contains privacy aspects and PReDUS highlight the words and the categories they belong to, similar to Figure 2b. The second US considers a borderline example, where the input US does not contain privacy aspects but it contains privacy-related words. PReDUS reports the missing privacy-related issues, and the user is made aware of the privacy words detected during the semantic analysis. Finally, the third US is obtained by performing two changes to the previous US, which modify the sense of the sentence also from the privacy point of view. PReDUS is capable of detecting the meaning of the modified sentence and highlights the privacy contents.

References

[1] D. Méndez Fernández, et al., Naming the pain in requirements engineering - contemporary problems, causes, and effects in practice, Empirical software engineering (2017) 2298–2338.

[2] S. H. Houmb, S. Islam, E. Knauss, J. Jürjens, K. Schneider, Eliciting security requirements and tracing them to design: an integration of common criteria, heuristics, and UMLsec, Requirements Engineering (2010) 63–93.

[3] P. Anthonysamy, A. Rashid, R. Chitchyan, Privacy requirements: Present future, in: Int. Conference on Software Engineering: SEIS track, 2017, pp. 13–22.

[4] F. Casillo, V. Deufemia, C. Gravino, Detecting privacy requirements from user stories with NLP transfer learning models, Information and Software Technology 146 (2022) 106853.

[5] L. Torrey, J. Shavlik, Transfer learning, in: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, IGI global, 2010, pp. 242–264.

[6] M. Cohn, User Stories Applied: For Agile Software Development, Addison Wesley, 2004.

[7] I. Sommerville, P. Sawyer, Requirements Engineering: A Good Practice Guide, Wiley, 1997.

[8] G. Lucassen, F. Dalpiaz, J. M. van der Werf, S. Brinkkemper, The use and effectiveness of user stories in practice, in: Req. Eng.: Found. for Soft. Quality, Springer, 2016, pp. 205–222.

[9] A. J. Gill, A. Vasalou, C. Papoutsi, A. N. Joinson, Privacy dictionary: A linguistic taxonomy of privacy for content analysis, in: Proceedings of Inter. Conference on Human Factors in Computing Systems (CHI), ACM, 2011, pp. 3227–3236.

[6] https://github.com/FrancescoCasillo/PReDUS-A-Privacy-Requirements-Detector-from-User-Stories