Automatic Recognition of Personal Data in Textual Documents

Đorđe Dragutinović*, Darko Čapko*, Marko Marković*, Stevan Gostojić*
* University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia
{djordje.dragutinovic, dcapko, markic, gostojic}@uns.ac.rs

***Abstract*—This paper presents an application used to recognize personal data in textual documents written in Serbian and English languages. Recognition is performed using SpaCy and Classla named entity recognition models. The results of personal data recognition are presented and analyzed.**

***Keywords*—personal data, named entity recognition, SpaCy, Classla**

I. INTRODUCTION

According to the General Data Protection Regulation (abbr. GDPR), "personal data" are any data that refer to a person whose identity has been or may be determined, directly or indirectly, on the basis of an identity mark or one or more features of his identity. Some of the personal data that are most often mentioned and used are name and surname, address, unique personal identification number and telephone number, but personal data also include a photograph, fingerprint, health and property status, religious and political affiliations and others [1].

The appearance of personal data is increasingly common in electronic documents which are stored in computer memory, and computer users are often not even aware of the large number of documents containing personal data. It is even possible to collect personal data without the knowledge or explicit permission of the data subject (for example, through surveillance camera footage or website cookies). With the development of the Internet, personal data have become much more accessible, since they are often kept in publicly available documents. This is the reason why personal data are most often mentioned in the context of their protection, so data anonymization and detection of documents containing personal data have to be often performed.

Because of that, it is necessary to solve the problem of recognizing personal data in documents by analysing electronic documents stored in computer memory and recognizing personal data contained in them. Recognition of personal data in documents could be done manually as well, but it would take a lot of time, so the goal is to develop a computer application that will give results that are similar with those obtained by manual annotation. The question we address here is whether it is possible to do that, and if yes, how successfully.

The structure of this paper is as follows: Section 2 presents an overview of the papers that solved similar problems. Section 3 describes details of methodology and models used for recognition of personal data. Implementation steps are listed and described in Section 4. Section 5 shows how the solution described in this paper can be used to recognize personal data in textual documents. The results of recognizing personal data in documents written in English and Serbian are listed and analyzed in Section 6. Finally, section 7 provides the summary of the work and describes its advantages and disadvantages, and gives a proposal on how the stated disadvantages can be corrected.

II. RELATED WORK

A large number of existing papers deal with the problem of recognizing personal data in textual documents. In paper [2], the authors present how existing, publicly available models for identifying personal data can be used in the analysis of biographies. The data set consisted of 249 manually annotated biographies, taken from the online encyclopedia "Wikipedia". Four existing, publicly available models were used to identify personal data. The evaluation of the used models was performed by calculating precision and recall, but the authors also defined a new way of evaluating, in order to take into account situations when personal information is a phrase consisting of several words, but the model successfully recognized only one part of that phrase. The results show that the precision of identifying personal data varies between 73% and 96%, and the recall varies between 64% and 97%, depending on the model used, and all the models have trouble recognizing named entities consisting of multiple words.

Paper [3] compares the success of existing models in recognizing personal data in newspaper articles written in English and Russian. A total of 7 models for personal data recognition were used, and four publicly available sets of annotated newspaper articles were used for testing the models. The evaluation of the used models was performed by calculating the F1-score, which represents the harmonic mean value of precision and recall. The results show that the models recognized the personal data much more precisely in articles written in English, due to the way the models were trained.

Paper [4] shows how existing models can be used to identify personal data in documents written in Malaysian. The data set was obtained by web scraping the articles from Malaysian newspaper portals, with a total of 3296 words in the data set, of which 229 represented personal

data. Two models were used to identify the personal data, and precision, recall and F1 value were used to show the success of the models. The analysis of the obtained results shows that both models gave extremely poor results in the recognition of personal data. It can be concluded that, due to the different morphology of English and Malaysian, the existing models cannot be successfully used for personal data recognition in documents written in Malaysian, unless additional training of the model is performed beforehand.

Most similar papers use existing, publicly available models to identify named entities, but in some of the papers the authors have tried to develop their own model [5, 6, 7]. However, a common drawback of all similar papers is that the models have to be included in source code by using some of the programming libraries in order to be used, which prevents users without programming knowledge from using them. The only solution made in the form of an application with a user interface is GATE [8]. However, the disadvantage of this solution is that only one document at a time can be selected for analysis.

III. METHODOLOGY

The idea of the solution is to find all textual documents saved in desired format, and load and analyze their content, in order to identify and classify personal data appearing in them. The easiest way to solve this problem is by using a method called Named Entity Recognition (abbr. NER). This method belongs to the field of artificial intelligence, and its goal is to recognize named entities in an unstructured text and perform their classification, in order to define which category they belong to. Since the documents written in Serbian and English were analyzed, before recognizing the named entities, it was necessary to recognize in which language was the analyzed document written.

Two free, publicly available models were used to identify personal data in textual documents:

• *SpaCy* - an open-source library used for advanced natural language processing. It was developed by the German software company "Explosion" and was written in the Python programming language. It is possible to perform alpha tokenization, part-of-speech tagging, text categorization, lemmatization, sentence segmentation, additional training of models and many other functions. The library is based on convolutional neural networks and the Thinc library for machine learning [9]. In the implemented application, this model was used to identify personal data in documents written in English.

• *Classla* (CLARIN Knowledge Center for South Slavic languages) - publicly available model for processing and analysing texts written in Slavic languages (Serbian, Croatian, Slovenian, Macedonian, and Bulgarian). It was developed by the Slovenian national institute CLARIN, and the existing Stanford NER model was used as a basis for creating this model, and the identification of named entities is done by using a combination of HMM (Hidden Markov Model) and MEMM (Maximum Entropy Markov Model) algorithms. It is possible to perform text tokenization, part-of-speech tagging, lemmatization, sentence segmentation, parsing of dependencies and recognition of named entities [10]. In the implemented application, this model was used to identify personal data in documents written in Serbian.

IV. IMPLEMENTATION

The following programming languages, environments and libraries were used in the implementation of the application for recognizing personal data in textual documents:

- *Python* - a high-level, general-purpose programming language.
- *PyCharm* - an integrated development environment for Python programming.
- *Tkinter* - a tool for creating applications with graphical user interface in Python.

The steps in the implementation of the solution are as follows:

- Load the content of the textual document (one or more) selected by the user
- Recognize the language in which the document was written
- Call the appropriate model for recognition of personal data
- Display recognized personal data

The first step in the solution is to load the content of the textual document. The application support working with 3 different types of documents (*.pdf*, *.docx* and *.txt*), so 3 different methods were used to allow loading of content. Loading content of .txt documents can be performed using *open()* function, which is included in every Python interpreter. On the other hand, the functionality of loading the content of documents saved in *.pdf* and *.docx* formats is not included in interpreters, so two libraries were used: *Apache Tika* and *docx2txt*.

If the content of the document is successfully loaded, it is necessary to recognize the language in which the document was written, in order to use a model which will give the best results for a detected language. To recognize the language the *langid* library was used, which offers the ability to recognize 97 different languages. Since only documents written in Serbian and English are taken into account, it is possible to recognize personal data only in such documents. All the other documents, which are found not to be written in one of these two languages, are neglected and recognition of personal data is not performed within them.

The loaded content is then passed as a parameter to the appropriate model, which will perform the recognition of personal data in it. After the recognition of personal data is performed, all recognized personal data are shown to the user in the main window of the application. For each recognized data, it is indicated to which group of personal data it belongs.

V. DEMONSTRATION

The main window of the implemented application consists of 3 units. The first unit is a large field, within which the results of personal data recognition are printed.

Below this field are 3 buttons, which offer the user ability to choose which types of documents will be analysed. At the bottom of the window there are 3 buttons, which offer the ability to select the location to be searched, the ability to select the concrete document to be analyzed and to start the process of personal data recognition. The appearance of the main window of implemented application is shown in Figure 1.

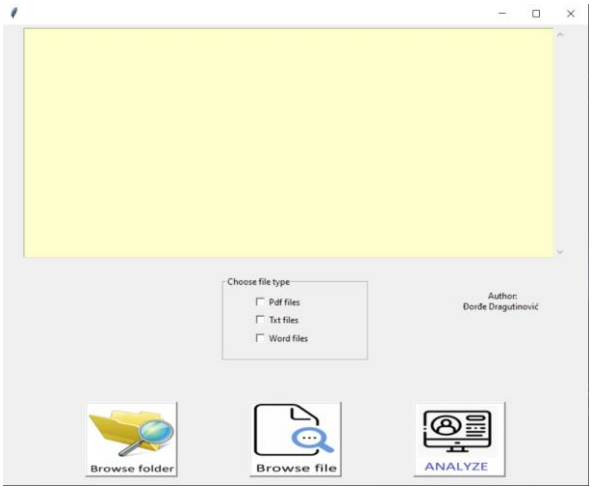

Figure 1. Main window

If the user has successfully selected the document or location he wants to analyze, clicking on the "Analyze" button performs loading of the document's content and recognition of personal data. The content of the analyzed file will be displayed in the upper half of the window, with all the words that do not represent personal data colored black, and personal data will be colored red or green, depending on the type. Names and surnames, which are personal data that most often appear in documents, will be colored red, while all other personal data will be colored green. A "bubble" appears in the main window, explaining to the user the way the data are colored. The appearance of the main window after recognizing personal data in the selected file is shown in Figure 2.

Figure 2. Main window after recognition of personal data

VI. RESULTS

Our solution was tested by performing automatic recognition of personal data in a set of documents, both on Serbian and English. The documents used to test the solution were previously manually annotated, in order to compare the model results with the results obtained by manual annotation. The results of personal data recognition, measured through precision, recall and F1 score, are shown in the Table 1.

	Precision	Recall	F1 score
English	0.87	0.92	0.89
Serbian	0.64	0.59	0.62

Table 1. Personal data recognition results

Since names and surnames are personal data that most often appear in textual documents, the results of recognizing this data have been singled out and are shown in Table 2.

	Precision	Recall	F1 score
English	0.85	0.92	0.88
Serbian	0.63	0.87	0.74

Table 2. Results of recognition of personal names and surnames

Obtained results show that the recognition of personal data is much more precisely performed in documents written in English. The reasons why the recognition in documents written in Serbian is less precise are the insufficient training of the Classla model and the complexity of Slavic languages compared to English. This problem can be solved by annotating manually large number of texts, documents and newspaper articles from various domains and fields, written in Serbian, which will be used for additional training of the model.

If we compare the results of recognition of personal names from Table 2 with the results of recognition of all personal data shown in Table 1, it can be concluded that, in documents written in English, the results of recognition of names largely coincide with the results of recognition of all entities. In documents written in Serbian, the recognition of names and surnames is much more successful than the average results of personal data recognition. This means that, in documents written in Serbian, the recognition of other types of personal data is quite imprecise, so additional training of model has to be performed.

VII. CONCLUSION

In this paper, we proposed the way to perform automatic recognition of personal data in textual documents, in order to find the documents in which personal data are mentioned and used. We have confirmed our hypothesis that it is possible to do so.

The biggest advantages of the implemented solution compared to similar solutions are ability to easily perform

recognition of personal data without the need to include the solution in source code, ability to easily select the location or document that needs to be analyzed, as well as the manner of displaying recognized personal data in the document. Instead of indicating only the position of the personal data within the text, the complete content of the document is displayed, with the personal data being colored separately.

Possible improvements of the implemented solution relate to the improvement of personal data recognition results. The current results are slightly worse compared to similar solutions, which is particularly evident when analyzing the results of recognizing personal data in textual documents written in Serbian. This problem can be solved by manually annotating a large number of documents and newspaper articles, which will then be used for additional model training.

REFERENCES

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [Online]

[2] S. Atdag and V. Labatut, "A comparison of named entity recognition tools applied to biographical texts", in Proc. of the 2nd International Conference on Systems and Computer Science, Villeneuve d'Ascq (FR), 2013

[3] S. Vychegzhanin and E. Kotelnikov, "Comparison of named entity recognition tools applied to news article", in Proc. of the 2019 Ivannikov Ispras Open Conference (ISPRAS)

[4] S. Sulaiman, R. Abdul Wahid, S. Sarkawi, and N. Omar, "Using Stanford NER and Illinois NER to detect Malay named entity recognition", in Proc. of the International Journal of Computer Theory and Engineering, Vol. 9, No. 2, April 2017

[5] E. Partalidou, E. Spyromitros-Xioufis, S. Doropoulos, S. Vologiannidis, and K. Diamantaras, "Design and implementation of an open source Greek POS Tagger and Entity Recognizer using SpaCy", in Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence, October 2019

[6] T. Ruokolainen, P. Kauppinen, M. Silfverberg, and K. Linden, "A Finnish news corpus for named entity recognition", *Springer.,* vol. 54, pp. 247-272, Aug. 2019.

[7] E. Milkov, R. Wang and W. Cohen, "Extracting personal names from email: applying named entity recognition to informal text", in Proc. of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language

[8] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "GATE: An architecture for development of robust HLT applications", in Proc. of the 40th Annual meeting of the Association for computational linguistics, Philadelphia, USA, July 2002

[9] SpaCy – Industrial-strength Natural Language Processing in Python [Online]. Available. https://spacy.io

[10] V. Batanović, N. Ljubešić, T. Samardžić, and M. Miličević Petrović, "Otvoreni resursi i tehnologije za obradu srpskog jezika", in Proc. of the Primena slobodnog softvera i otvorenog hardvera 2020 (PSSOH 2020), Beograd, Srbija