Using AI and crowdsourcing in digitisation and processing of archival materials

Stančić, Hrvoje; Seljan, Sanja; Ivanjko, Tomislav

ИСТОРИЯ СОВРЕМЕННОСТИ

Информационные ресурсы, методы и исследовательские практики в России и за рубежом

Доклады
Международной
научно-практической конференции

Москва, 28–29 ноября 2019 г.

Москва
2019

УДК 94(470)(042.3)
ББК 63.3(0)ф.я43
И90

Под общей редакцией *Е.П. Малышевой*
Составитель *Л.Д. Шаповалова*

Сборник издается при финансовой поддержке Фонда «История Отечества»

Доклады публикуются в авторской редакции

ISBN 978-5-7281-2577-8

Using AI and crowdsourcing in digitisation and processing of archival materials

H. Stančić
S. Selja
T. Ivanjko
University of Zagreb

Abstract

Digitisation of archival materials is a lengthy process in which the act of digitisation is usually the shortest and the most straightforward one. It is the processing of the digitised materials that take a lot of time and effort. It is valuable to have digitised materials available online, even for the reason of not having to travel to the physical location of the originals, but the tendency of extracting additional value from the materials increases. To that end, the archival materials can be processed by new techniques such as artificial intelligence (AI) and crowdsourcing. The authors use the example of digitisation of the minutes from the Faculty (of Humanities and Social Sciences) Council meetings dating from 1913 to 1996 and the application of AI to improve the OCR results and NER for semantic enrichment as well the example of digitisation of the food rationing cards, used between 1941 and 1945 in Zagreb, Croatia and the application of AI and crowdsourcing to data extraction, analysis and visualisation.

Keywords: archives, archival materials, digitisation, artificial intelligence, name entity recognition, crowdsourcing.

Использование искусственного интеллекта и краудсорсинга при оцифровке и обработке архивных материалов

Х. Станчич
С. Селян
Т. Иванько
Университет Загреба

Аннотация

Оцифровка архивных материалов – длительный процесс, в котором выполнение непосредственно оцифровки обычно является самым коротким и простым. Именно обработка оцифрованных материалов занимает

больше всего времени и сил. Полезно иметь оцифрованные материалы, доступные в интернете, хотя бы по той причине, что в таком случае отсутствует необходимость приезжать к физическому местоположению оригиналов, но одновременно усиливается тенденция извлечения дополнительной ценности из материалов. С этой целью архивные материалы могут быть обработаны с помощью новых методов, таких как искусственный интеллект (ИИ) и краудсорсинг. Авторы приводят пример оцифровки протоколов заседаний Совета факультета гуманитарных и социальных наук с 1913 по 1996 г. и применения ИИ для улучшения результатов OCR и NER с целью семантического обогащения. Другой пример – оцифровка продовольственных карточек, использовавшихся в период с 1941 по 1945 г. в Загребе (Хорватия), и применение ИИ и краудсорсинга для извлечения, анализа и визуализации данных.

Ключевые слова: архивы, архивные документы, оцифровка, искусственный интеллект, распознавание лица, краудсорсинг.

Introduction

Many heritage institutions are faced with challenges of the modern information society. Emerging media technologies that change the information landscape, together with new user habits and expectations have started to redesign the relationships between users and institutions. A great number of archival materials are being increasingly digitized, and the traditional practices of archiving are being transformed. The first statement of the ICA constitution defines the mission of the archives to "...constitute the memory of nations and societies, shape their identity, and are a cornerstone of the information society"[1]. In order to fulfil such a mission, archives need to explore contemporary practices. In this context, one of the main challenges is to open as much of the archival materials as possible to online communities in ways that enable democracy, accountability and good governance as the basic principles of archival practice.

The first step of that process is to digitise archival materials and it is usually the easiest and the most straightforward step. Once digitised, archival materials should be further processed. The process of Optical Character Recognition (OCR) may produce optimal results but usually for the more modern, typewritten or printed materials. In the case of older typewritten texts, as detailed later, the OCR may not be efficient enough so that the AI-based solutions need to be used in order to improve the recognition results.

The handwritten materials may be recognised using Intelligent Character Recognition (ICR) or Handwritten Text Recognition

(HCR) software, which is far less efficient than the OCR when applied to the printed texts and, if combined with the machine learning (ML), need to be trained to efficiently recognize handwriting of one writer. The process needs to be repeated, and the model trained, for the next writer. For example, the Transkribus platform[2] suggests that for the training of a model between 5,000 and 15,000 words (around 25-75 pp.) of transcribed material are required[3]. The neural networks will learn by comparing the images of the text with the transcribed material and produce the trained model that can be used to efficiently recognize further scanned images of the handwritten material of the same writer.

After the successful text recognition, archival materials can be further processed to get semantically enriched. A possible way to extract specific information from corpora is to use Name/Named Entity Recognition (NER) techniques. Through the process of annotation, the text is enriched by additional information which enables semantic information research and additional data analysis.

However, in some cases the machine recognition is simply not producing good-enough results. The solution to that can be to take a crowdsourcing approach and ask the users of archival materials to participate in the process of transcription.

Next, the crowdsourcing approach will be explained followed by the discussion about the possibilities of NER techniques in the context of digital collection building. The examples of application of those two approaches and application of the AI techniques to the process of text recognition in two digitisation projects will be shown in order to argue for integrational strategy to building of digital archival collections.

Crowdsourcing in the archive

A growing number of institutions in the heritage sector started to investigate the possibilities of communicating their digitised collections with the public and seizing the opportunities that arise from digitisation by applying different approaches to user engagement. The efforts in completing such tasks are often gathered under the notion of *citizen science*, where projects enlist the help of many volunteers to solve challenging scientific research problems. There are a number of websites, such as Zooniverse (URL: https://www.zooniverse.org/), Citizen Archivist Dashboard (URL: https://www.archives.gov/citizen-archivist) or Smithsonian Digital Volunteers (URL: https://transcription.si.edu) that enable people to take part in real cutting-edge research in many fields across the sciences, humanities, and more. The combination of human computation and sociality has shown to be highly effective, not only in accomplishing the original scientific

objectives that were outsourced to the crowd, but also in yielding unanticipated discoveries initiated by the members of the community[4]. In the heritage sector, such efforts are often gathered under the notion of *crowdsourcing,* a term derived from the word *outsourcing*, and it represents the act of a company or institution taking over a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call[5]. The field of crowdsourcing has become increasingly popular in recent years, and a number of recent works have investigated different applications of crowdsourcing systems and initiatives[6] showing that the field has influenced the ways in which products and services are invented, produced, funded, marketed, distributed[7]. Many projects that include crowdsourcing often use different gamified approaches to engage the community. Such approaches apply the idea of gamification – the introduction of gaming elements in otherwise non-gaming contexts[8]. Enhancing information systems with game elements has shown to have a positive impact on the effectiveness and efficiency of employees, and gamification has attracted significant attention from the fields of computer science, informatics, human-computer interaction, game studies and psychology[9]. Applications span from education and training to health, self-management, innovation, employee engagement, heritage, crowdsourcing, civic engagement, and marketing. Gamification is nowadays an established practice and industry segment, by some estimates poised to grow to over US$ 11 bn by 2020[10].

Many crowdsourcing projects have shown benefits from complementary perspectives: productivity and engagement. From a productivity standpoint, engaging in citizen science projects can achieve goals that the institution would not have the resources (temporal, financial, or staff-related) to achieve and it can also add value to digital collections. This means that a strategy for building digital collections should take into consideration the planning for user metadata ingest and quality control. From an engagement standpoint, crowdsourcing projects can actively engage the community in communicating with the institution and its systems and collections by building up a relationship with the community through building mutual trust and encouraging loyalty to the institution. Once properly engaged, and the example of the Transcribe Bentham project which has 10,000 registered users[11] shows that it can be done, the community can submit transcripts at the fast rate. The transcriptions need to be checked, corrected where needed and further processed. The named entity recognition is one possible approach to further processing aiming at enabling semantic information research.

Named Entity Recognition

The Name/Named Entity Recognition (NER), also known as entity identification or entity extraction, represents a subfield of information extraction. This method is widely used in many fields of artificial intelligence (AI), such as Natural Language Processing (NLP) and Machine Learning (ML). NER is used in order to extract relevant information from the large set of data, which is afterwards analysed and used for various purposes.

The NER techniques have recently been used in various industries and activities, such as in tourism[12], law[13], for marketing purposes, opinion detection, customer segmentation, in insurance organizations, and financial institutions such as banks[14]. The NER techniques are also applied to financial documentation and publicly available non-financial dataset to extract specific information[15] for detecting trends, possible threats, marketing effects or sentiments. NER is widely used in unstructured or semi-structured text mining process[16].

The NER process mainly consists of two steps – detection of entities in the text and classification by the type of entity – but also of discovering relationships among entities. In the detection process, problems of segmentation can appear (e.g. the "National Bank of Croatia" which is a single name, as opposed to "Croatia", which is a location) which might have significant impact on the subsequent classification.

In the NER process, previously defined categories can be used, e.g. *Person*, *Location*, *Organization*, *Date*, *Time*, etc. The NER techniques enable detection of the existing predefined categories from the scanned or transcribed documents. That way, it is possible to automatically identify, e.g. locations, cities, or universities mentioned in the digitised materials.

The NER systems can be created by using grammar-based rules, statistical models or ML techniques. The language-specific grammar-based rules offer the best precision but require long-term specialist work of language engineers. Statistical NER techniques require large amount of manually annotated data used for training, while in the ML process, supervised and semi-supervised approaches are used.

The NER technique uses simple block of texts. E.g. the sentence:

Prof. Novak from the University of Zagreb held a visiting lecture at University of Graz on the International Cooperation Day in 1989.

can be transformed into the annotated block of texts with the following sample of categories:

[Prof.] $_{\text{Title}}$ [Novak] $_{\text{Person}}$ from the [University of [Zagreb] $_{\text{City}}$] $_{\text{Institution}}$ held a [visiting lecture] $_{\text{Activity}}$ at [University of [Graz] $_{\text{City}}$] $_{\text{Institution}}$, [Austria] $_{\text{Country}}$ on the [International Cooperation Day] $_{\text{Event}}$ in [1989] $_{\text{Time}}$.

However, NER techniques may fail in different situations, when entities are not detected, partially detected (e.g. "Cooperation Day" instead of the "International Cooperation Day") or detected with wrong assignment (e.g. "Ford" as personal name instead of company), with smaller or larger scope (e.g. identifying "James Madison" as a personal name instead of "James Madison University" as an institution). Also, NER systems may be challenged by the domain coverage, where a NER system developed for one domain does not perform well in other domains. Nevertheless, despite possible shortcomings, NER can be quite useful.

Next, using the example of two projects it will be explained how AI for OCR, NER and crowdsourcing can be seamlessly integrated in the context of digitisation and processing of archival materials.

Integration of AI, NER and crowdsourcing in the digitisation projects

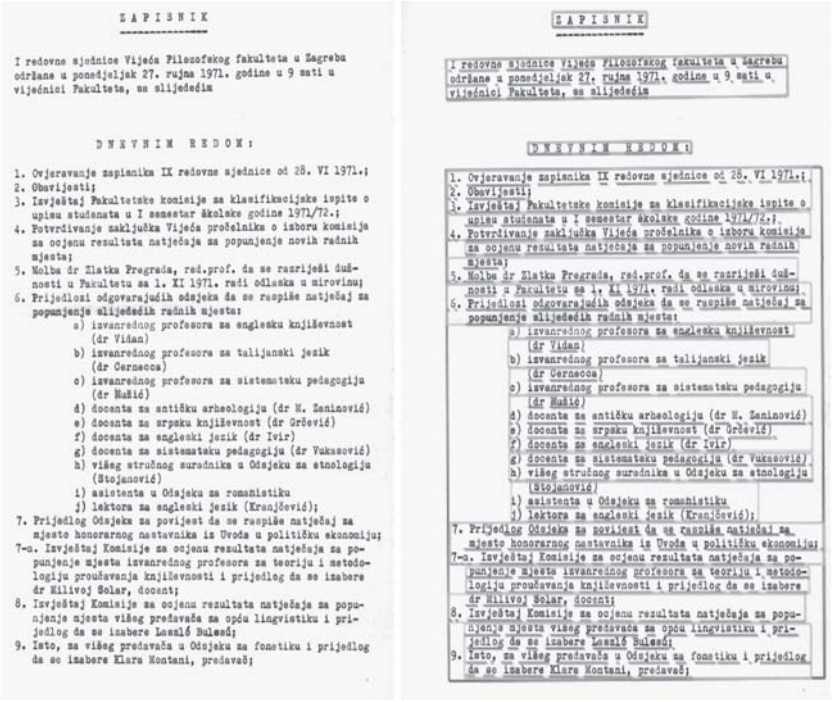
Digitisation of the minutes from the Faculty Council meetings

The first example is the project "Digitisation of the archival materials from the Archives of the Faculty of Humanities and Social Sciences, University of Zagreb and the development of digital humanities (DAFF)" (2017-). The project aim is to digitise minutes from the Faculty Council meetings from 1874 until the digital era. The minutes are partly handwritten and partly typewritten. The handwritten materials are written by different hands thus making the training set increasingly complex. The typewritten texts are often the 2nd or the 3rd copies on the thin 40g paper. That fact impeded the results of the several tested commercial and open source OCR solutions. Therefore, an alternative approach was needed. It was found in the Google Cloud's Vision AI (URL: https://cloud.google.com/vision/). With the technical support of the Bonsai.AI (URL: https://bonsai.tech/), a Croatian-based AI company, it was possible to achieve better text recognition results than using standalone OCR software.

Google OCR is part of the Google Cloud Vision API services. It supports two types of recognition – text detection in images (e.g. text on a billboard that is part of an image) and document text detection. The Google OCR is available as a web-service through the API calls. The POST requests are sent in the following form:

```
{  "requests": [{
      "features": [{ "type": "DOCUMENT_TEXT_DETECTION" }],
      "image": { "content": "(data from FF-Z-1971-72-1971-09-27-R01-0001.jpg)" }
}]}
```


where the content is the data from the *base64-encoded-image* (in this example a .jpg image of the 1st page of the minutes of the 1st regular Faculty Council meeting in the academic year 1971-72 held on 27 November 1971). The system segments the document text into text blocks, paragraphs, sentences, words and characters (Fig. 1).



ZAPISNIK

I redovne sjednice Vijeća Filozofskog fakulteta u Zagrebu održane u ponedjeljak 27. rujna 1971. godine u 9 sati u vijećnici Fakulteta, sa slijedećim

DNEVNIM REDOM:

1. Ovjeravanje zapisnika IX redovne sjednice od 28. VI 1971.;
2. Obavijesti;
3. Izvještaj Fakultetske komisije za klasifikacijske ispite o upisu studenata u I semestar školske godine 1971/72.;
4. Potvrđivanje zaključka Vijeća pročelnika o izboru komisija za ocjenu resultata natječaja za popunjenje novih radnih mjesta;
5. Molba dr Zlatka Pregrada, red.prof. da se razriješi dužnosti u Fakultetu sa 1. XI 1971. radi odlaska u mirovinu;
6. Prijedlozi odgovarajućih odsjeka da se raspiše natječaj za popunjenje slijedećih radnih mjesta:
   a) izvanrednog profesora za englesku književnost (dr Vidan)
   b) izvanrednog profesora za talijanski jezik (dr Cernecca)
   c) izvanrednog profesora za sistematsku pedagogiju (dr Mužić)
   d) docenta za antičku arheologiju (dr M. Zaninović)
   e) docenta za srpsku književnost (dr Grčević)
   f) docenta za engleski jezik (dr Ivir)
   g) docenta za sistematsku pedagogiju (dr Vuksović)
   h) višeg stručnog suradnika u Odsjeku za etnologija (Stojanović)
   i) asistenta u Odsjeku za romanistiku
   j) lektora za engleski jezik (Kranjčević);
7. Prijedlog Odsjeka za povijest da se raspiše natječaj za mjesto honorarnog nastavnika iz Uvoda u političku ekonomiju;
7-a. Izvještaj Komisije za ocjenu resultata natječaja za popunjenje mjesta izvanrednog profesora za teoriju i metodologiju proučavanja književnosti i prijedlog da se izabere dr Milivoj Solar, docent;
8. Izvještaj Komisije za ocjenu resultata natječaja za popunjenje mjesta višeg predavača za opću lingvistiku i prijedlog da se izabere Laszló Bulešić;
9. Isto, za višeg predavača u Odsjeku za fonetiku i prijedlog da se izabere Klara Montani, predavač;

Fig. 1. Comparison of the scanned image and the image segmented into text blocks

The system recognises the segmented parts and provides feedback. Each recognised word is described with its coordinates (four coordinates representing four corners of the segmented block) as shown in the following excerpt of the JSON code («\n» in the *description* part denotes beginning of the new line; "..." denotes that part of the code is clipped):

The original scanned pages of the minutes, along with the results of the recognition, are combined in PDF files. The files retain the original "look and feel" while the additional text layer with the recognised text is used for text search and further analysis possibilities. The files are ingested in the Indigo – a hybrid repository platform based on semantic technologies created by the Croatian company ArhivPRO.


```
{
  "textAnnotations": [{
      "boundingPoly": {
        "vertices": [
          {"x": 301,"y": 202},
          {"x": 2244,"y": 202},
          {"x": 2244,"y": 3002},
          {"x": 301,"y": 3002}
        ]},
      "description": "ZAPISNIK\n-\nI redovne sjednice Vijeća Filozofskog
      fakulteta u Zagrebu\nodržane u ponedjeljak 27. rujna 1971. godine u 9
      sati u\nkijesnici Fakulteta, sa slijedećim\n",
    ...
      "locale": "hr"
    },
    ...
    { "boundingPoly": {
        "vertices": [
        {"x": 313,"y": 403},
        {"x": 326,"y": 403},
        {"x": 326,"y": 449},
        {"x": 313,"y": 449}
        ]},
      "description": "I"},
    { "boundingPoly": {
        "vertices": [
          {"x": 365,"y": 402},
          {"x": 572,"y": 400},
          {"x": 572,"y": 447},
          {"x": 365,"y": 449}
        ]},
      "description": "redovne"
    },
    { "boundingPoly": {
        "vertices": [
          {"x": 610,"y": 399},
          {"x": 854,"y": 397},
          {"x": 854,"y": 444},
          {"x": 610,"y": 446}
        ]},
      "description": "sjednice"
    },
    ...
]}
```


The minutes are described according to the ISAD (G). It was possible apply the NER approach and enable semantic enrichment. For example, it is possible to search for a person and narrow the search down only to the results when the person was acting as a committee chair, or as a dean, or is mentioned as a professor etc. The places are also annotated and with the further advance of the project it might be possible to conduct a network analysis of e.g. places where the visiting professors were coming from in certain periods.

Digitisation of the food rationing cards

The project "Digitisation and computer processing of archival materials" (2019-) aims at digitisation of the food rationing cards used between 1941 and 1945 in Zagreb, Croatia. Three institutions cooperate on the project – The State Archives in Zagreb, Faculty of Humanities and Social Sciences, University of Zagreb and Bons.AI.

The rationing cards could also be viewed upon as a form of census. There are around half a million of them in the collection. They contain names of all members of a household at a certain address, relations between the members (e.g. son, daughter, maid etc.), their year of birth, occupation, place of work, salary, amount of groceries present at home, address of the ration distribution centre etc. The information is older than 70 years so they can be freely open to the public.

The rationing cards are printed forms, in several variations, mostly filled in by hand (Fig. 2), and are in some cases typewritten. This poses a problem for the optical text recognition. Therefore, the mixed AI and crowdsourcing approach will be taken. The coordinates of the handwritten information will be identified using GC Vision AI and the extracted information will be ingested into a crowdsourcing platform

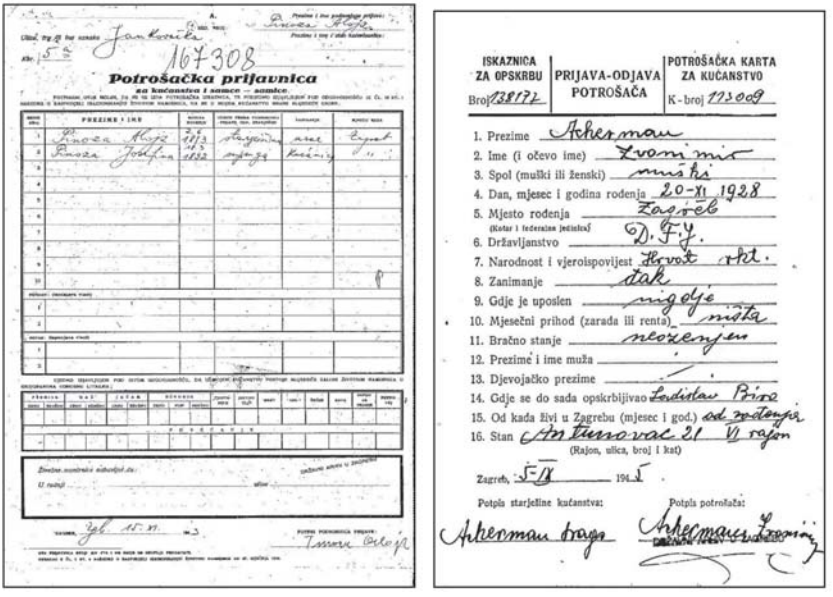


Fig. 2. Examples of rationing cards

for the users to transcribe. After that, the data will be NER analysed, visualised etc. The project has just started, and the first results are expected in late 2020.

Conclusion

The examples of the digitisation projects show that the digitisation and processing of archival materials can be far from easy. The materials need to be selected for digitisation, digitised, processed before they undergo the text recognition process, either using an OCR software or applying an AI approach involving pattern recognition and machine learning. Upon recognition, the materials need to be further processed by the NER techniques in order to be further semantically enriched. The NER techniques are not flawless but can offer valuable insights for detection of the specific data. On the other hand, if the text recognition does not prove to be efficient, the user community can be engaged. Crowdsourcing is not just a method of getting things done by the users, but an approach that can engage them to contribute, collaborate and co-create. Many successful projects in the heritage sector, often gathered under the notion of citizen science, have shown that crowdsourcing can be a viable solution to a number of tasks and can help shifting the users' focus from merely consuming digital collections to collaborating in their development.

To conclude, embracing the possibilities offered by the AI, NER and crowdsourcing can improve digitisation and processing of archival materials. Therefore, they should be integrated in the institutional digitisation strategies.

References

[1] ICA: International Council on Archives (2018), Mission, aim and objectives, available at: https://www.ica.org/en/mission-aim-and-objectives [Accessed: 10.30.2019].

[2] READ: Recognition and Enrichment of Archival Documents (2019), How to Train A Handwritten Text Recognition Model in Transkribus. v.1.8.0 (24.10.2019.), available at: https://transkribus.eu/wiki/images/3/34/HowToTranscribe_Train_A_Model.pdf [Accessed: 10. 31.2019].

[3] Ibid.

[4] Tinati R. et al. (2017), An investigation of player motivations in Eyewire, a gamified citizen science project. *Computers in Human Behavior*, 73, 527-540.

[5] Howe J. (2006), *The rise of crowdsourcing,* available at: https://www.wired.com/2006/06/crowds/ [Accessed: 10.30.2019].

[6] Brabham D.C. (2013), *Crowdsourcing*. MIT Press. Cambridge, MA, USA; Geiger D., Schader M. (2014), Personalized task recommendation in crowdsourcing information systems – Current state of the art. *Decision Support Systems*, 65, 3-16; Zuchowski O., Posegga O., Schlagwein D., Fischbach K. (2016), Internal crowdsourcing: Conceptual framework, structured review and research agenda. *Journal of Information Technology*, 31, 166-184.

[7] Tapscott D., Williams A.D. (2011), *MacroWikinomics: Rebooting business and the world*. Atlantic Books, London.

[8] Deterding S. et al. (2011), From game design elements to game fulness: defining gamification. In: Lugmayr, Artur et al. (eds). *Proceedings of the 15th international academic Mind Trek conference: envisioning future media environments*. New York: ACM, 9-15.

[9] Stieglitz S. et al. (2017), *Gamification: using game elements in serious contexts*. Cham: Springer.

[10] Nacke L.E., Deterding S. (2017), Editorial: The maturing of gamification research. *Computers in Human Behaviour*, 71, 450-454.

[11] H2020 Project READ (Recognition and Enrichment of Archival Documents). 2016–2019 (2019), available at: https://www.academia.edu/22653102/H2020_Project_READ _Recognition_and_Enrichment_of_Archival_Documents_-_2016-2019 [Accessed: 10.31.2019].

[12] Chantrapornchai C., Tunsakul A. (2019), Information Extraction based on Named Entity for Tourism Corpus. *International Joint Conference on Computer Science and Software Engineering (JCSSE)*. doi: 10.1109/JCSSE.2019.8864166.

[13] Dozier C., Kondadadi R., Light M., Vachher A., Veeramachaneni S., Wudali R. (2010), Named Entity Recognition and Resolution in Legal Text. In: Francesconi E., Montemagni S., Peters W., Tiscornia D. (eds.), *Semantic Processing of Legal Texts. Lecture Notes in Computer Science,* vol 6036. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-12837-0_2.

[14] Marrara S., Pejić Bach M., Seljan S., Topalovic A. (2019), FinTech and SMEs: The Italian Case. FinTech as a Disruptive Technology for Financial Institutions, Rafay, Abdul (ed.). Hershey, Pennsylvania: IGI Global, 42-60. doi:10.4018/978-1-5225-7805-5.ch002; Pejić Bach M., Krstić Ž., Seljan S. (2019). Big data text mining in the financial sector. *Expert Systems in Finance: Smart Financial Applications in Big Data Environments*. Metawa, Noura; Elhoseny, Mohamed; Hassanien, Aboul Ella; Hassan, M. Kabir (eds.). London: Taylor & Francis Group: Routledge, 80-96. doi:10.4324/9780429024061; Pejić Bach M., Krstić Ž., Seljan S., Turulja L. (2019b), Text Mining for Big Data Analysis in Financial Sector: A Literature Review. *Sustainability,* 11, 5; 1-27. doi: 10. 3390 /su11051277.

[15] Alvarado J.C.S., Verspoor K., Baldwing T. (2015), Domain Adaptation of Named Entity Recognition to Support Credit Risk Assessment. *Proceedings of Australasian Language Technology Workshop,* 84-90.

[16] Saju J.C., Shaja A.S. (2017), A Survey on Efficient Extraction of Named Entities from New Domains Using Big Data Analytics. *2nd International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM*), 170-175.