Proceedings of the of the 17th Conference on Computer Science and Intelligence Systems pp. 715–719
DOI: 10.15439/2022F49
ISSN 2300-5963 ACSIS, Vol. 30

Modelling an IT solution to anonymise selected data processed in digital documents

Barbara Probierz*, Tomasz Jach*, Jan Kozak*, Radosław Pacud† and Tomasz Turek‡
* Department of Machine Learning,
University of Economics in Katowice,
1 Maja, 40-287 Katowice, Poland
Email: barbara.probierz,tomasz.jach,jan.kozak@ue.katowice.pl
† Faculty of Finance,
University of Economics in Katowice,
1 Maja, 40-287 Katowice, Poland
Email: radoslaw.pacud@ue.katowice.pl
‡ Faculty of Management,
Częstochowa University of Technology,
al. Armii Krajowej 19 B, 42-201 Częstochowa, Poland
Email: tomasz.turek@pcz.pl

***Abstract*—Allowing access to real legal documents is an important element both for the development of science and the judiciary. On the other hand, protecting information about citizens or organizations, that appear in these documents, is crucial and required by law. Therefore, before the documents are distributed, the data anonymisation process should be carried out. Unfortunately, there is no perfect tool that can automatically anonymise documents in such a way, that the main concept of the document is preserved; especially in the case of documents written in inflectional language. The aim of this article is to show how important (and at the same time how difficult) is the task to identify personal or corporate data of a client, as well as other related personal data in documents that are subject to legal protection. We conducted research aimed at assessing the usefulness of IT techniques as well as decision rules and patterns in the anonymisation of legal documents. A set of real legal documents written in Polish was used for the research in which we identified selected types of data that need to be anonymised. Eventually, the obtained results were assessed by field experts. Additionally, in order to verify the effectiveness of the proposed solution, we conducted research on a set of 50,000 false identities with names, company names, addresses and other confidential information. The collection was created using Fake Name Generator[1]. The obtained results from both experiments confirmed that the solutions we proposed is accurate even in the case of real legal documents.**

I. Introduction and Related works

THE PROCESS of anonymising documents is important to protect individuals and institutions from the illegal dissemination of information [1]. However, in the case of legal documents, this is a very difficult process due to unstructured content of the documents [2]. In addition, the provisions of law, i.e. the GDPR [3], will force institutions to appropriately transform their current IT systems in the field of personal data processing. Due to the large variety of document structures and the lack of a uniform system applicable to all judicial organizations, there is a high risk of privacy breach. Due to the variety of document structures and the lack of a uniform system applicable to all judicial organizations, there is a high risk of privacy breach. For this reason, the anonymisation process is often widely regarded as an expensive and inefficient process [4]. Additionally, due to the low effectiveness of the available data anonymisation tools, the anonymisation process in law firms is most often performed manually by trained employees [5].

In this article, we will look at the theoretical assumptions of a basic IT tool designed to identify selected types of data in text documents. On the other hand, the research goal is to check and evaluate the usefulness of IT techniques as well as decision rules and patterns in the anonymisation of legal documents. In the case of legal documents, data identifying individuals or organizations must be removed, anonymised or pseudonymised from the digital document following the end of the legal service. Authors are convinced that if such data are not manually obliterated - either manually or using a computer-assisted method - any lawyer should delete all documents of clients from law firm's digital media and computers in a time demanded by national regulation of legal occupation or implicit period of time assumed by the professional need (processing by the purpose). By all above mentioned regulations personal data and other types must be deleted or pseudonymised by legal tech. To achieve this goal, a team of coauthors is working on a project designed to build such a technology[2].

The first step in anonymising text documents is the ex-

[1]https://www.fakenamegenerator.com/

[2]Project developed by Infojura Sp. z o. o. (KRS 0000502117) carried out in the period from 01/01/2022 to 31/12/2023 under the name "Technologies for Automatic anonymisation of Personal Data" in digital documents. It is financed with the support of EU as part of the Regional Operational Program of the Silesian Voivodeship for 2014-2020 (European Regional Development Fund). Action 1.2. Research, development and innovation in enterprises.

traction of information that directly identifies individuals or organizations. They are referred to as Named Entities (NE) [6]. Most often, NE are classified into predefined semantic categories, i.e. first name, last name, organization name or location [7], [8]. Natural language processing (NLP) methods are used to automatically recognize NE [9]. For this purpose, Named Entity Recognition (NER) [10] was created. It is the process of locating information in the text, which then becomes a specific category. In addition to the basic categories, i.e. the name of an organization or person, NER has added categories that define time and numerical expressions, such as monetary values [11].

In the process of identifying named entities, the language in which the text document is written is very important. In the case of languages with an extensive morphological structure, the processing of text documents and the extraction of information is very difficult [12]. Therefore, Graliński et al [13] presented a formalism for the rule-based NER. Their research focuses on the use of NER for inflectional languages (especially for Polish and Czech) and the translation of Named Entities. Researchers developed two applications that could be used for machine anonymisation and machine translation. Similarly, J. Pisowski [14] created the formalism of recognizing NE from Polish texts and manually created a set of NER rules for the Polish language.

To assist researchers in sharing raw textual data, Kleinberg et al. [15] proposed the NETANOS anonymisation system that identifies and modifies named entities (e.g. people, locations, times, dates). Bayesian tests showed that NETANOS anonymisation was practically equivalent to human anonymisation. The authors only used NER to detect personal data. Using this method, they hypothesized that the ability to discover the original meaning of anonymised data could ultimately be similar to that of humans if the training data set is large enough.

Unfortunately, the NER in the field of law, despite its importance, is not a well-researched area. Many current approaches use different techniques and classification methods on different datasets [16]. Additionally, most of the obtained anonymisation results are only assessed by experts. For this reason, it is not possible to properly compare the results. However, the proposed solutions make a significant contribution and are the basis for further research. It should be remembered that an additional problem in the field of law is the natural language of legal documents and the different legal provisions in individual countries.

C. Dozier et al. [17] were one of the first to conduct research based on Named Entity Recognition in the legal field. Using the example of American case law, testimonies, pleadings and other legal documents, the authors analyzed the NER. C. Cardellino et al. [18] have developed a tool to identify, classify and link legal NE. The authors focused on four different levels of detail, one of which was NER. They used a Stanford NER [19] support vector machine and a neural network in the learning algorithm. Elena Leitner et al. [20] presented the problem of fine-grained recognition of entities in legal documents. The authors developed a data set consisting of decisions of German courts, in which the source texts were manually annotated with 19 semantic classes. Then all classes were automatically generalized to seven classes (person, location, organization, legal norm, individual regulation, court order and legal literature). The results obtained show that there is no universal model with the best understanding of all classes.

II. Research methodology

PII (Personally Identifiable Information) data detection algorithms have to be fine-tuned for each case. It is virtually impossible to obtain a high accuracy with low noise on the same pass of detection. Whereas, in the case of data related to digits, even though a number is a valid KRS/NIP (tax identification numbers) value, only semantic context might indicate whether it is a true PII or just a coincidence.

Therefore, in the case of the analysis of specific documents and adapted to a specific language (legal documents prepared in Polish), we proposed to use a combination of this methods:

- Dictionary search. A relatively easy search withing the known dictionary. The example of this method is used in our experiments where searching for first names (both female and male). The names were taken from official PESEL database[3]. However convenient, this list has a lot of potential to give many false positives, as the PESEL number (unique person identification number) was given to a lot of non-Polish citizens; often with short names being homonyms of common Polish words (like “Na”).
- Decision rules and patterns. Using this approach, one is using regular expressions to describe the patterns of potential PII data. For instance, a NIP number is usually consisted of 10 digits divided by hyphens.
- Special type of validation for self-checking numbers. NIP number are self-checking numbers, as they use a Luhn algorithm [21], as well as additional constrains for numbers being valid. Checking this constrains is a great way to decrease greatly the number of false positives.
- Validation using external services is often used for IBAN numbers or KRS numbers in Poland. KRS number is a 10 digit number with no additional constrains, being just a next element in a single sequence of all entities in Poland. Thus, an external validator is used to make sure that the 10 digit number is an actual and up to date KRS for a valid company.
- Heuristic search using some known patterns like "sp z o.o." for a private limited company. Due to law regulations in Poland, some equities need to include mandatory information in company name.

III. Computational experiments

The research objective of this paper was to investigate the applicability of decision rule and pattern search methods in legal document anonymisation. Therefore, in order to test the

[3]https://dane.gov.pl/pl/dataset/1667,lista-imion-wystepujacych-w-rejestrze-pesel-osoby-zyjace

proposed approach, we conducted two types of experimental studies. In the first one, we wanted to test the efficiency of the proposed solution for large amounts of data – artificial data generated for this purpose. In the second one, we performed the actual anonymisation of legal documents, and their evaluation was done by domain experts.

The first research attempt, related to achieving the stated goal of anonymising selected data in legal documents, was to test the algorithm on artificial data. In this case we used Fake Name Generator[4] service and generated 50000 fake identities with names, company names, addresses and other sensitive information. Each line in this set should contain exactly one of name, email, phone and company name. The perfect scenario will have a mean value of detection equal to 1 for each column. The data is adapted to the Polish language and contains on average 73 words and 422 characters per line. The data preparation was followed by the application of the proposed solution to each case. The results of the experiments are presented by a box plot (Fig. 1) in which, for each PII, the minimum, 1st, 2nd (median) and 3rd quantile and maximum values are given – often these values are repeated, which is why whole box plots do not appear in the figure.

As it can be seen, in the case of the name, too many values are detected (the maximum found is 8, the median is 2, and the expected value, which is 1, is also the 1st quantile). This is because the first name often appears in the address (e.g. street name established by a famous person); in one line, in addition to the correctly detected name, the one from the address is also recognised. The ideal situation is in the case of e-mail address, where all cases were correctly detected; so was the case of phone number (there, in one case, apart from the correct number, also the false positive is detected). For the KRS number, as it is just a sequential number, false positives were found in different parts of data, as it was the case in first name. In addition to the correctly detected KRS number also NIP number is included – this explains the very low detectability of NIP number. In these experiments, once detected, the element could not be recognised again. We proposed a different approach in later experiments on real data.

The worst situation appears for the company name, which was basically not detected. This is because in our heuristic approach we require the presence of certain keywords. The generator in question was not characterised by the presence of appropriate prefixes in the name. However, in the case of legal documents, such full names are always used.

In order to assess the achievement of the final goal, the proposed method has been applied to a real-life study of a law firm operating in the territory of Poland. Fifteen legal documents written in Polish and concerning different legal scopes and cases were analysed. All documents were originally saved in DOCX format, from which, in order to apply the model, they were converted into text format. The analysis was based on the documents of different length and number of words per line (see Tab. I). It can be observed that the low

[4]https://www.fakenamegenerator.com

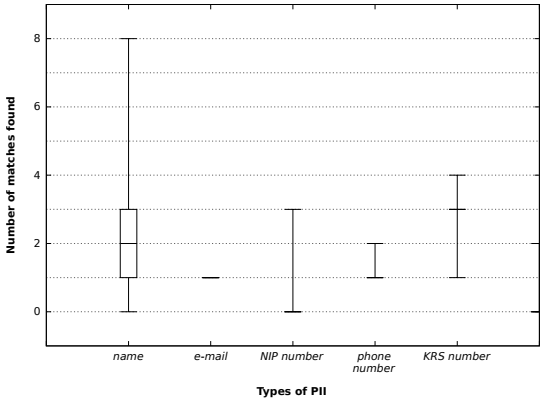

Fig. 1. Minimum, maximum, and quantiles for the prepared data

TABLE I
FEATURES OF THE DOCUMENTS ANALYSED

Legal document	# lines	# words	$\frac{\#\text{words}}{\#\text{lines}}$
x_1	104	500	4.81
x_2	500	3730	7.46
x_3	121	617	5.10
x_4	335	1800	5.37
x_5	162	1788	11.04
x_6	664	4291	6.46
x_7	59	294	4.98
x_8	399	2906	7.28
x_9	50	130	2.60
x_{10}	57	201	3.53
x_{11}	180	532	2.96
x_{12}	30	74	2.47
x_{13}	139	840	6.04
x_{14}	29	123	4.24
x_{15}	135	1474	10.92
sum	2964	19300	6.51
avg.	198	1287	6.50
median	135	617	4.57

number of words per line is related to the short information provided in the legal document, such as the invoice number or the service name. Therefore, the table I presents the average number of words per line.

In the next step, the fields indicated for anonymisation (with the reason for anonymising an element) were verified by specialists. They verified the PII described in Section II and each time described three known elements of the classification quality assessment:

- TP – true positives, i.e. elements that really should be anonymised for the reason given;
- FP – false positive, i.e. elements for which the indicated reason for anonymisation is not appropriate.
- FN – false negative, i.e. items that should have been anonymised for a given reason but were not.

The evaluation of TN (true negative), i.e. elements that should not be anonymised, was omitted and indeed this was not done. TN was not evaluated because, naturally, all non-annotated words mean true negative for this problem. De facto this is the number of all words in the document, except those in TP, FP and FN.

Analysing the exact anonymisation result given in Tab. II

TABLE II
RESULTS OF THE ANONYMISATION OF LEGAL DOCUMENTS

Legal	name			e-mail			NIP number			phone number			KRS number			company name		
document	TP	FP	FN	TP	FP	FN	TP	FP	FN	TP	FP	FN	TP	FP	FN	TP	FP	FN
x_1	5	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
x_2	8	10	5	0	0	0	0	0	0	0	0	0	0	0	0	2	13	0
x_3	12	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x_4	0	3	0	0	0	0	3	0	0	0	3	0	3	3	0	5	0	0
x_5	2	1	0	0	0	0	0	0	0	0	0	0	0	1	0	7	10	0
x_6	3	6	0	3	0	1	3	0	0	0	0	0	4	4	0	6	0	0
x_7	0	1	0	0	0	0	3	0	0	0	0	0	4	3	0	4	1	0
x_8	13	2	0	0	0	0	3	0	0	0	0	0	5	5	0	5	1	0
x_9	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
x_{10}	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
x_{11}	4	0	1	0	0	0	6	0	0	0	0	0	3	6	0	10	1	3
x_{12}	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x_{13}	1	0	2	0	0	0	2	0	0	0	0	0	2	2	0	2	0	0
x_{14}	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	2	0	0
x_{15}	2	0	1	0	0	0	2	0	0	0	0	0	1	2	0	1	0	1

and reading the actual document, it can be seen that in the case of PII–name there are many FP, i.e. false positive indications for anonymisation. This is mainly due to street names (consistent with first name) and first names that are also names of months (e.g. 'Maja'). FN-related errors, on the other hand, occur in the case of an atypical linguistic variant of a name.

E-mail addresses only appeared in one document and were mostly well recognised. The error associated with the address not being found was due to a transcription error (the domain extension was single character). Thus, future consideration could be given to analysing email addresses not only according to the rules, but also assuming human error in the transcription.

The situation looks very good in the case of PII–NIP number, which is a tax identification number that was correctly recognised in all documents. This is largely made possible by the Luhn algorithm. The phone number, on the other hand, did not actually appear in any document, although three times the algorithm incorrectly indicated that the found string of digits was a phone number.

An interesting situation concerns the PII–KRS number, i.e. the national court register number. All KRS numbers were detected (true positive – TP), but also other numbers in the document were incorrectly indicated as KRS. This was most often the case with the NIP number, but there were also other strings of numbers which did not relate to the KRS (it should be noted, however, that in further work, each of the elements indicated in the experiments as KRS should be anonymised anyway – whether as NIP number or national identifier of a person).

Big problems arise in the case of company names. For the analysed data, the algorithm often incorrectly indicated data for anonymisation although it was not necessary (false positive – FP). There were also situations where the company name was not detected at all. However, to sum up the work of the algorithm, in the vast majority of the cases, the algorithm detected company names that should be anonymised.

At the same time, it should be noted that in this type of documents it is a much bigger mistake to omit an element to be anonymised (i.e. those indicated in FN) than to anonymise it incorrectly (in our case FP). In addition, often elements that showed FP for one of the methods were in fact anonymised anyway for another reason. Therefore, in Tab. III, the summary results are presented, in which the reason for anonymisation is not indicated, but only the information whether a given word should be anonymised or not.

TABLE III
RESULTS OF THE ANONYMISATION OF LEGAL DOCUMENTS IN TERMS OF TP, FP, FN AND IN TERMS OF MEASURES TO ASSESS CLASSIFICATION QUALITY: PRECISION, RECALL AND F–SCORE

Legal document	TP	FP	FN	precision	recall	F–score
x_1	6	1	1	0.8571	0.8571	0.8571
x_2	10	23	5	0.3030	0.6667	0.4167
x_3	12	1	1	0.9231	0.9231	0.9231
x_4	11	6	0	0.6471	1.0000	0.7857
x_5	9	12	0	0.4286	1.0000	0.6000
x_6	19	6	1	0.7600	0.9500	0.8444
x_7	11	2	0	0.8462	1.0000	0.9167
x_8	26	3	0	0.8966	1.0000	0.9455
x_9	8	0	0	1.0000	1.0000	1.0000
x_{10}	2	2	2	0.5000	0.5000	0.5000
x_{11}	23	1	4	0.9583	0.8519	0.9020
x_{12}	3	0	0	1.0000	1.0000	1.0000
x_{13}	7	0	2	1.0000	0.7778	0.8750
x_{14}	4	0	0	1.0000	1.0000	1.0000
x_{15}	6	0	2	1.0000	0.7500	0.8571
sum	157	57	18	0.7336	0.8971	0.8072
avg.	10	4	1	0.8080	0.8851	0.8448
median	9	1	1	0.8966	0.9500	0.9225

With these values, it can be indicated that in all analysed documents 185 words should have been anonymised (out of a total of 19300 words), with 168 words actually detected, 17 words not anonymised (although they should have been), and 57 words incorrectly indicated as requiring anonymisation. This gives a median of 10 words per document (requiring anonymisation) while omitting 1 word per document and indicating 1 additional word incorrectly. This allows us to assess the effectiveness of the solution.

With the values in Tab III, it is possible to calculate measures of classification evaluation, such as $precision$ (eq. (1)), $recall$ (eq. (2)) and $F - score$ (eq. (3)). These are important

measures, where the $precision$ determines with what certainty we can trust the classifier that a given element (word) should actually be anonymised, while $recall$ determines how many elements (words) that should be anonymised in the document have been indicated as anonymised, and $F - score$ is the harmonic mean of $precision$ and $recall$.

$$precision = \frac{TP}{TP + FP} \tag{1}$$

$$recall = \frac{TP}{TP + FN} \tag{2}$$

$$F - score = \frac{TP}{TP + 0.5 \cdot (FP + FN)} \tag{3}$$

As mentioned earlier, from the point of view of the anonymisation of legal documents, it is more important to detect the elements that should be anonymised. Therefore, when analysing the results, the recall measure should be optimised first. The results related to the quality assessment of the anonymisation work are presented in Tab. III. In the case of the proposed solution, our method achieves a recall of 89.71% (in terms of the sum of all documents) and 95.00% in terms of the median recall calculated separately for each legal document.

Of course, it is important to note that all the results reported in this section relate to the analysis of only 6 types of PII: name, e-mail address, tax identification number (NIP number), phone number, national court register number (KRS number) and company name. Other data that should actually be anonymised have not been analysed by us at this stage of the project – the proposed solution. For subsequent work, a more sophisticated use machine learning and natural language processing methods is required.

IV. Conclusions and future works

It should be emphasised that the research objective was to analyse a limited range of data – only a selected type of data was anonymised. Thus, we wanted to assess the applicability of IT techniques and decision rules and patterns in document anonymisation. The conducted experiments confirm that the solutions we have applied allow us to obtain good results – this is particularly evident in the case of real legal documents. Therefore, in the future, in addition to improving the rules proposed so far, we believe that it is worth to develop the modelling of machine learning on bigger groups of real legal documents that can be used as learning data in supervised machine learning models.

Future work may include developing a multi-criteria model for multi-step analysis. It is possible to initially verify data on the basis of unit names, and then carry out a thorough analysis on the basis of the content only for selected data. Such a solution is possible with the use of machine learning methods and natural language processing techniques. However, it should be remembered that the task of creating a universal system that would recognize all classes is very difficult, and the effectiveness of anonymisation largely depends on the structure of the documents that could be referred to given types of legal documents.

References

[1] P. Štarchoň and T. Pikulík, "Gdpr principles in data protection encourage pseudonymization through most popular and full-personalized devices-mobile phones," *Procedia Computer Science*, vol. 151, pp. 303–312, 2019.

[2] M. Mozes and B. Kleinberg, "No intruder, no validity: Evaluation criteria for privacy-preserving text anonymization," *arXiv preprint arXiv:2103.09263*, 2021.

[3] P. Regulation, "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) (text with eea relevance)," *Regulation (eu)*, vol. 679, p. 2016, 2016.

[4] I. Glaser, T. Schamberger, and F. Matthes, "Anonymization of german legal court rulings," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 2021, pp. 205–209.

[5] G. M. Csányi, D. Nagy, R. Vági, J. P. Vadász, and T. Orosz, "Challenges and open problems of legal document anonymization," *Symmetry*, vol. 13, no. 8, p. 1490, 2021.

[6] B. Mohit, "Named entity recognition," in *Natural language processing of semitic languages*. Springer, 2014, pp. 221–245.

[7] T. H. Cao, T. M. Tang, and C. K. Chau, "Text clustering with named entities: a model, experimentation and realization," in *Data mining: Foundations and intelligent paradigms*. Springer, 2012, pp. 267–287.

[8] R. Grishman and B. M. Sundheim, "Message understanding conference-6: A brief history," in *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996.

[9] H. Vico and D. Calegari, "Software architecture for document anonymization," *Electronic Notes in Theoretical Computer Science*, vol. 314, pp. 83–100, 2015.

[10] B. M. Sundheim, "Overview of results of the muc-6 evaluation," in *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*, 1995.

[11] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[12] O. Kabasakal and A. Mutlu, "Named entity recognition in turkish bank documents," *Kocaeli Journal of Science and Engineering*, vol. 4, no. 2, pp. 86–92, 2021.

[13] F. Graliński, K. Jassem, M. Marcińczuk, and P. Wawrzyniak, "Named entity recognition in machine anonymization," *Recent Advances in Intelligent Information Systems*, pp. 247–260, 2009.

[14] J. Piskorski, "Named-entity recognition for polish with sprout," in *Intelligent Media Technology for Communicative Intelligence*. Springer, 2004, pp. 122–133.

[15] B. Kleinberg and M. Mozes, "Web-based text anonymization with node. js: Introducing netanos (named entity-based text anonymization for open science)," *Journal of Open Source Software*, vol. 2, no. 14, p. 293, 2017.

[16] D. Reynders, "Digitalising justice systems to bring out the best in justice," *Eucrim: the European Criminal Law Associations' fórum*, no. 4, pp. 236–237, 2021.

[17] C. Dozier, R. Kondadadi, M. Light, A. Vachher, S. Veeramachaneni, and R. Wudali, "Named entity recognition and resolution in legal text," in *Semantic Processing of Legal Texts*. Springer, 2010, pp. 27–43.

[18] C. Cardellino, M. Teruel, L. A. Alemany, and S. Villata, "A low-cost, high-coverage legal named entity recognizer, classifier and linker," in *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, 2017, pp. 9–18.

[19] J. R. Finkel, T. Grenager, and C. D. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)*, 2005, pp. 363–370.

[20] E. Leitner, G. Rehm, and J. Moreno-Schneider, "Fine-grained named entity recognition in legal documents," in *International Conference on Semantic Systems*. Springer, 2019, pp. 272–287.

[21] H. P. Luhn, "Computer for verifying numbers," *US Patent*, vol. 2, no. 950, p. 048, 1960.