Using Natural Language Processing to Detect Privacy Violations in Online Contracts

Paulo Silva, Carolina Gonçalves, Carolina Godinho, Nuno Antunes, Marilia Curado
CISUC, Department of Informatics Engineering, University of Coimbra, Portugal
pmgsilva@dei.uc.pt,[mariapg,godinho]@student.dei.uc.pt,[nmsa,marilia]@dei.uc.pt

ABSTRACT

As information systems deal with contracts and documents in essential services, there is a lack of mechanisms to help organizations in protecting the involved data subjects. In this paper, we evaluate the use of named entity recognition as a way to identify, monitor and validate personally identifiable information. In our experiments, we use three of the most well-known Natural Language Processing tools (NLTK, Stanford CoreNLP, and spaCy). First, the effectiveness of the tools is evaluated in a generic dataset. Then, the tools are applied in datasets built based on contracts that contain personally identifiable information. The results show that models' performance was highly positive in accurately classifying both the generic and the contracts' data. Furthermore, we discuss how our proposal can effectively act as a Privacy Enhancing Technology.

CCS CONCEPTS

• **Security and privacy → Domain-specific security and privacy architectures**; • **Computing methodologies** → *Information extraction*; Feature selection;

KEYWORDS

Privacy Violations, Online Contracts, Natural Language Processing, Named Entity Recognition, Personally Identifiable Information

ACM Reference Format:
Paulo Silva, Carolina Gonçalves, Carolina Godinho, Nuno Antunes, Marilia Curado. 2020. Using Natural Language Processing to, Detect Privacy Violations in Online Contracts. In *The 35th ACM/SIGAPP Symposium on Applied Computing (SAC '20), March 30-April 3, 2020, Brno, Czech Republic.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3341105.3375774

1 INTRODUCTION

Recent data breaches and privacy scandals have likely triggered discussion, more specific policy-making and further research within the privacy area. In Europe, civil society, academia, industry, and policymakers are driven by GDPR-compliance [19], as well as its practical and legal effects.

Certain public or private organizations are legally bound to release contractual information. Such contracts contain specific information (e.g., persons' names, addresses, financial or employment information or dates) about not only organizations but individuals as well. To comply with regulations and efficiently increase privacy assurances, it is necessary to develop mechanisms that can not only provide such privacy assurances but also increased automation and reliability. The automated monitoring of *Personally Identifiable Information* (PII) can effectively be measured in terms of reliability while being properly automated at the same time. The usage of *Natural Language Processing* (NLP) and *Named Entity Recognition* (NER) are the ideal candidates to monitor and detect privacy violations not only in such contracts but also in a broader scope.

Literature [6, 13] provides sufficient guidance on choosing the right NLP tools and mechanisms for specific tasks. The most commonly mentioned NLP tools are the *Natural Language Toolkit* (NLTK) [20], Stanford CoreNLP [3] and spaCy [5]. Moreover, NER systems are equally analysed [1, 2, 18] regarding their overall accuracy performance, recognition of named entities in tweets or biomedical data, respectively. Nevertheless, there is insufficient work relating NER and PII in its broad-spectrum, which encompasses many different kinds of personal information.

In this work, we argue that NLP, and NER can be a very adequate *Privacy Enhancing Technology* (PET) when applied in privacy-preserving data analysis as this avoids the usage of dictionary approaches and the involvement of human operators. Supporting our claims are the results of the experiments we conducted. First, we start by analysing three NLP tools regarding their characteristics and capabilities. Further on, we used a generic publicly available dataset, partitioned it, and assessed the performance of the NLP tools in multiple dataset sizes. Then, we used datasets that contained publicly available PII (e.g., names, addresses, contract numbers or other related types), namely contracts. After manually labelling the entities and training the models, it was possible to observe that the $F_1 score$ of our models was approximately 90% in the best cases. Therefore, the results show how NLP and NER models can be applicable as a PET.

2 BACKGROUND

NLP is a branch of *Artificial Intelligence* (AI) that helps computers understand, interpret and manipulate human language. NER is one of NLP's sub-tasks that seeks to find and classify named entities present in a text into specific and pre-defined categories [9]. Those categories can be people's names, addresses, states, countries, money, organisations, laws, date, etc. To classify those entities, different notation schemes can be used, and their main difference is the amount of information that can be encoded. Each NLP tool has a NER that accepts a certain notation. Therefore, performing NER in different NLP tools may lead to different NER performances. Also, a NER system designed within a tool for one project may execute differently in another project or not do the task at all [11].

SAC '20, March 30-April 3, 2020, Brno, Czech Republic

ACM ISBN 978-1-4503-6866-7/20/03.
https://doi.org/10.1145/3341105.3375774

The *Machine Learning* (ML) domain covers several topics, algorithms, and applications. It is a branch of AI research field that allows the software to learn and make predictions by finding patterns or any other kind of statistical regularity. The NLP tools used in this work employ different machine learning algorithms to classify data. In the case of NLTK [20], a Naive Bayes classifier [7] and *Hidden Markov Models* (HMM) [8] are applied. Naive Bayes classifiers share a common principle: every pair of features being classified is independent of each other [7]. For Stanford CoreNLP [3], *Conditional Random Fields* (CRFs) are applied. CRFs are probabilistic models that perform segmentation and labelling of sequential data [8]. SpaCy [5] uses *Convolutional Neural Networks* (CNNs) with pre-trained word vectors to train its models. CNNs are *Neural Networks* (NNs) used primarily to classify images, cluster images by similarity, and perform object recognition within scenes [12]. Nevertheless, CNNs are not limited to image recognition they also have been applied directly to text analytics.

3 EXPERIMENTAL METHODOLOGY

In the first stage, we gathered generic data that was already tagged with named entities. Additionally, we partitioned the dataset in different sizes from 2.5% to 100% of the original size. In the second stage, we focused on publicly available contracts containing PII. In the first stage, a dataset with 1.354.149 tokens (based on the Groningen Meaning Bank data [16]) was retrieved from Kaggle [10]. In the second stage, a dataset with 19.838 tokens was created as a fusion between contracts available in online sites [14, 17] and other contracts from the U.S *Department of Defense* (DoD) [4]. For all datasets, there was a 70% and 30% proportion for training and validation, respectively.

To analyse and evaluate the performance of the NLP tools and its respective NER system, we identified the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) derived from the model classification. Having such information allowed to calculate the $F_1 score$: the harmonic mean of the Precision and Recall of the models. The time spent on training, as well as the number of iterations needed were also recorded.

4 EXPERIMENTAL RESULTS

4.1 General Model

The results obtained with the previously labelled general model are illustrated in Figure 1. NLTK obtained a $F_1 score$ of 0,47 using the smallest portion of the dataset (2,5%). On the other hand, Stanford CoreNLP and spaCy achieved approximately 0,65. The full size dataset (100%) got the best scores: NLTK achieved approximately 0,67, while Stanford CoreNLP and spaCy obtained 0,84 and 0,86, respectively. Nevertheless, there is no significant difference between the 20%-sized dataset and the larger ones that follow. The $F_1 score$ difference between the 20%-sized dataset and the full dataset is between 0,03 and 0,05.

In our experiments, the worst performance was NLTK. On the other hand, although Stanford CoreNLP and spaCy achieved similar results, the best performance was seen in spaCy with a small margin. The results indicate that without any comprehensive tuning of the model training settings, spaCy provides the best results for F1-Score. Additionally, we observed than training the models for less than 500 iterations provides similar results and requires much less training time.

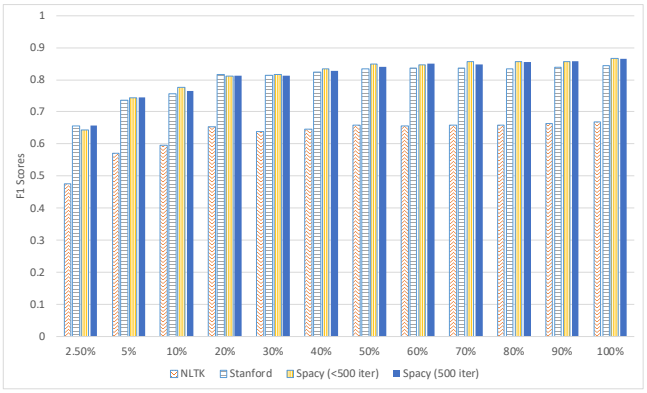

Figure 1: F_1 Scores (NLTK, Stanford CoreNLP and spaCy)

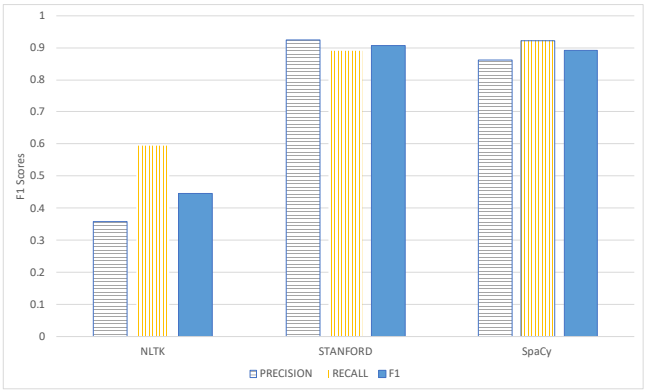

Figure 2: F_1 Scores, Precision and Recall values (NLTK, Stanford CoreNLP and spaCy)

4.2 Contract-based Model

Figure 2 shows the precision, recall, and $F_1 scores$ obtained while evaluating the models that were created from manually-labelled contracts. It is possible to observe that NLTK's best results were approximately 0,45. On the other hand, Stanford CoreNLP and spaCy reached very similar values (approximately 0,90). Moreover, the difference between these two is 0.01, being Stanford CoreNLP the one with better results in this case.

The less positive aspect is the time necessary for manually hand-labelling the data fed to the models. Each document took, on average, 4,75 hours of work. The measure indicates the time spent by one person labelling the entities in the referred datasets. Afterwards, each document was reviewed by at least one other person for consistency purposes.

4.3 Lessons Learned

Overall, the models performed similarly regardless of having generic or PII-specific content for training purposes. By knowing the behaviour of the machine learning algorithms behind such systems,

one could say that this should be the expected outcome. However, the size of the sample may not fully define the model's behaviour with other kinds of data. Nevertheless, with approximately 20 hours of manual labelling, it is possible to create a model that can identify entities such as person, city, title, employment details, and others.

There was a total of 47959 sentences (the same as 1.354.149 tokens) readily available and labelled. Nevertheless, we concluded that 20% of the total dataset size is sufficient to provide results that are very similar to the ones obtained using the full-sized dataset. The proportion of 20% is equivalent to 9590 total sentences, and the $F_1 score$ variation is minimal. For the second stage of the experiments, there were 1150 total sentences, which is approximately 10% of what we found to be the ideal size. However, this was due to the effort required to manually label data. Based on the entities we defined as PII, we used 68% of them during the labelling process. To counter this issue, we are currently assessing the possibility of using Mostly AI [15] to generate synthetic data or to recur to online annotation services.

5 APPLICABILITY AS A PRIVACY ENHANCING TECHNOLOGY

General data validation is enforced in a wide variety of services and fields. With our approach, systems could be able to not only validate data types and formats but also the contents. Systems managing textual data inputs would be able to distinguish if the inputs match the actual description. This kind of monitoring can be applied in scenarios such as data exchanges between systems and/or users, documents or databases, depending on the context and the privacy implications. This would allow the system to warn data owners or systems administrators about the PII at stake.

Permission-based systems would be able to map and verify if the actual data matches the textual description of the respective permissions granted. In systems where there no such textual descriptions of the permissions, it is possible to directly map the permission type, to the PII type, thus allowing permission verification on a higher level.

All the data is processed and then discarded every time the system runs. This is because it uses ephemeral storage and it does not communicate with any other service for PII data exchange. To generate trustworthiness and to show compliance with *General Data Privacy Regulation* (GDPR) and other privacy-related regulations, we intend to release the system as open-source.

6 CONCLUDING REMARKS

In this work we evaluated the effectiveness of three different NLP tools and their NER sub-tasks in discovering PII and demonstrated how the proposed approach can effectively be used as a Privacy Enhancing Technology. We developed an experimental setup where different machine learning models were trained and evaluated with generic datasets as well as contracts with PII. The positive results of our proposal are verified by two main NLP tools (Stanford CoreNLP and spaCy). Although the trade-off between effort and benefits is not yet fully optimized, we show how this approach can reliably automate the monitoring of PII in different scenarios.

ACKNOWLEDGMENTS

The work presented in this paper was partially carried out in the scope of the PoSeID-on project - Protection and control of Secured Information by means of a privacy enhanced Dashboard, Grant Agreement Number: 786713. H2020-DS- 2016-2017/ DS-08-2017.

REFERENCES

[1] Ritter A., Clar. S, Mausam, and Etzioni O. 2011. Named Entity Recognition in Tweets: An Experimental Study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Edinburgh, Scotland, UK., 1524–1534.

[2] Vlachos A. 2007. Evaluating and combining and biomedical named entity recognition systems. In *Biological, translational, and clinical language processing*. Association for Computational Linguistics, Prague, Czech Republic, 199–200.

[3] Manning C., Surdeanu M., Bauer J., Finkel J., Bethard S., and McClosky D. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Baltimore, Maryland, 55–60. https://doi.org/10.3115/v1/P14-5010

[4] U.S Department Of Defense. 2019. Official website for U.S. Department of Defense. Retrieved August 27, 2019 from https://www.defense.gov/Newsroom/Contracts

[5] ExplosionAI. 2019. spaCy - Industrial-Strength Natural Language Processing. Retrieved August 27, 2019 from https://spacy.io

[6] Omran F. and Treude C. 2017. Choosing an NLP Library for Analyzing Software Documentation: A Systematic Literature Review and a Series of Experiments. In *Proceedings of the 14th International Conference on Mining Software Repositories (MSR '17)*. IEEE Press, Piscataway, NJ, USA, 187–197. https://doi.org/10.1109/MSR.2017.42

[7] Chen J., Huang H., Tian S., and Qu Y. 2009. Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications* 36, 3 (2009), 5432–5435.

[8] Lafferty J., McCallum A., and Pereira F. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 282–289. http://dl.acm.org/citation.cfm?id=645530.655813

[9] Li J., Sun A., Han J., , and Li C. 2018. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2145–2158.

[10] Kaggle. 2019. An online community of data scientists and machine learners. Retrieved August 27, 2019 from https://www.kaggle.com

[11] Ratinov L. and Roth D. 2009. Design Challenges and Misconceptions in Named Entity Recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 147–155. http://dl.acm.org/citation.cfm?id=1596374.1596399

[12] Zhang L. and Suganthan P. 2016. A survey of randomized algorithms for training neural networks. *Information Sciences* 364 (2016), 146–155.

[13] Christopher D Manning, Christopher D Manning, and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT press, London, UK.

[14] Metrolink. 2019. Metrolink is Southern California's premier regional passenger rail system serving over 55 stations across the region. Retrieved August 27, 2019 from https://www.metrolinktrains.com/globalassets/about/contracts/may-26-2019/contract-no.-sp452-16-conformed-contract-fully-executed.pdf

[15] MostlyAI. 2019. Creating AI-generated Synthetic Data. Retrieved August 27, 2019 from https://mostly.ai

[16] University of Groningen. 2019. Groningen Meaning Bank. Retrieved August 27, 2019 from https://gmb.let.rug.nl

[17] Texas Department of Information Resources. 2019. Our mission is to provide technology leadership, technology solutions. Retrieved August 27, 2019 from https://dir.texas.gov/View-Search/Contracts-Detail.aspx?contractnumber=DIR-TSO-4101

[18] Jiang R., Banchs R., and Li H. 2016. Evaluating and Combining Name Entity Recognition Systems. In *Proceedings of the Sixth Named Entity Workshop*. Association for Computational Linguistics, Berlin, Germany, 21–27. https://doi.org/10.18653/v1/W16-2703

[19] General Data Protection Regulation. 2016. 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. *Official Journal of the European Union (OJ)* 59, 1-88 (2016), 294.

[20] Bird S., Klein E., and Loper E. 2009. *Natural Language Processing with Python*. O'Reilly Media, Boston, USA.