Redaction detection in publicly published government papers

Student: Ammar Alhashmi - ammar.alhashmi@student.uva.nl
Supervisor: Maarten Marx - M.J.Marx@uva.nl

ACM Reference Format:
Student: Ammar Alhashmi - ammar.alhashmi@student.uva.nl and Supervisor: Maarten Marx - M.J.Marx@uva.nl. 2022. Redaction detection in publicly published government papers. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 ABSTRACT

Not all documents are made up entirely of language-based information. Non-linguistic features in a document, such as redacted regions that are usually present in documents as geometrical shapes, may contain useful information but are difficult to decipher by OCR or other state of the art techniques. As a result, a specific solution to recognize and comprehend such aspects is required, and that is the why behind this research paper. This research will investigate the use of state of the art techniques (such deep learning and machine learning) in order to detect redacted regions in PDF documents. To be more clear, an attempt will be made to detect the regions in PDF documents that hide/whitewash text or information. Deep-learning and machine learning models will be used to address this problem. The expected output from this research will be presented by showing the accuracy of the segmentation metrics.

2 KEYWORDS

Image segmentation, Deep-Learning, Redaction, Machine Learning, Image processing.

3 INTRODUCTION

In the netherlands, the Open Government Act requires the government to actively release information (known as WOO). This means that the government makes information public so that everyone can see it. Forinstance, the released information could be on regulations, laws, meeting and / or decision documents. Some government data isn't available to the general public. Because it contains sensitive information, for instance. The interested party must submit a WOO request to acquire that information. Previously, this was known as a WOB request. The Government Information (Public Access) Act has been replaced by the Open Government Act (WOO) [5]. In case the requests of information is approved, documents regarding the requested information are disclosed to the public. However, there are some absolute exceptions such as: the unity of the Crown and personal data. Thus, if the documents that would be published contain personal data, such data would be redacted / whitewashed [1]. This research paper is focused on the detection of the redacted regions in the released WOO documents.

In literature, several approaches have been used to detect, segment or extract features and information from PDFs such as region detection, image segmentation, colors detection and shape detection. For example, [22] presents a representative local region detector based on color-contrastpixel ranking and [21] proposes extracting more information and features from PDFs by converting PDFs to HTML form.

Furthermore, different research papers proposed several ways to information segment images to detect certain objects. Such as using a Convolutional Neural Network (CNN) structure To segment images. [11] considered approaches such as 2D CNN and a fully convolutional network (FCN) to segment medical images. FCN showed a superior accuracy.

What most of the research papers have in common is that they used deep learning on non textual images most of the time such as day to day object or medical images such as mitochondria images. What was insufficient from the literature is conducting features extraction and / or segmentation on textual documents to detect redaction, a dataset with textual images that contain redaction would look very different and have different features than the other datasets that were used in all the mentioned research paper. Most of the research papers used fixed datasets such as the COCO dataset. There are no fixed datasets that contain textual images with redacted regions. In this paper, an attempt was made to successfully detect redacted regions in WOO documents by applying deep learning and machine learning techniques. In other words, different approaches will be applied to detect the deleted or hidden text (whitewashed text) in PDF documents and the best approach will be chosen. The models performance will be checked by comparing the scores of segmentation metrics. The running time will be considered since the models must go through large PDF documents.

The research question that this research is trying to answer is: **RQ:** To What extent can state-of-the-art algorithms in image segmentation and classification improve the segmentation metrics (such as Panoptic Quality) [13] when detecting and calculating the percentage of redacted regions in WOO pages? **SQ1:** How do baseline CNN models based on [20] preform on WOO pages when detecting redacted regions while using a single source data (such PDFs from a specific municipality/ministry)? **SQ 2:** How can this performance be improved by enriching the data by using an annotation tool [19] and Convolutional filters with a traditional machine learning algorithm such as rainforest? **SQ 3:**How do these models perform when applied to WOO pages from different sources?

4 RELATED WORK

This section will discuss the related work. Firs the redaction (whitewashing) will be discussed and after that, the related literature that

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

relates to object, edge and shape detection will be discussed, since the discussed methods can help detect redaction.

4.1 What is Redaction and how it is done?

Several papers discussed redaction when it comes to sensitive information. Redaction is the process of removing visible information from a document [6]. This research investigate redaction in government documents that were made public. The information that was redaction was about private data. Private data is any personally identifying information that could be used to determine the producer's identity, the identity or personal information of individuals known to the producer (e.g., friends, relatives, and clients), or is linked to a private record (e.g., medical, employment, and education). Social security numbers, credit card numbers, bank records, medical records, employment information, education records, passwords and cryptographic keys, and local and online account records are examples of this. The reason why redaction in this case is important is because there are three primary hazards to digital heritage if collecting institutions do not improve their methods for discovering, identifying, and redacting sensitive content. First, collecting institutions may be viewed as untrustworthy actors incapable of properly caring for digital collections. As a result, digital content creators may be hesitant to hand over their works to institutions for long-term storage. Second, if processing costs are unreasonably high, institutions are likely to acquire fewer collections than they would otherwise [2]. Now more on the redaction process. When the information that needs to be redacted is identified. There are two approaches mentioned by [7]. The first one is called conventional redaction: The naive approach is to decompress the image, redact a region encompassing the offending pixels (e.g., replace with black or background values), and possibly recompress the image if the pixels in the image contain identifying information in the form of burned-in text that needs to be removed to protect privacy. The second one is called Block selective redaction: Because only a tiny number of pixel regions must be redacted to remove the identifying information, and the JPEG process divides the image into MCUs of small numbers of 8 x 8 blocks, it is clear that only those blocks that are affected must be edited, while the rest can be left alone. This method entirely eliminates any loss in areas that do not intersect the redacted areas. Every original block that is completely contained in a redacted zone can then be replaced with a new block. After the redaction process was over, [19] suggested a metric to calculate the accuracy of the process. Mean Intersection Over Union (mIoU) was suggested which is an evaluation metric used to measure the accuracy of an object detector on a particular dataset. the mIoU can be calculated using the following formula. $\sum(tp/(tp + fp + fn))$.

It is worth mentioning that the authors of this paper [19] labeled the ground truth image dataset by hand to ensure accuracy.

4.2 Segmentation in Deep-learning and traditional machine learning techniques

This section will discuss the state-of-the-art segmentation approaches in deep-learning and traditional machine learning respectively.

4.2.1 Deep-learning . Deep learning algorithms such as CNNs (Convolutional Neural Network) are effective not only in traditional tasks such as object classification but also in other related tasks such as image segmentation. CNN-based algorithms, which are one of the most widely used approaches in computer vision, have undergone a number of minor tweaks to improve their performance in segmentation tasks [18].

But first, let's have a deeper understanding of the conventional Convolutional Neural Network (CNN), CNN is a type of neural network that consists of a stack of layers, each of which performs a different function, such as convolution, pooling, loss computation, and so on. The output of the previous layer is sent into each intermediate layer. The first layer is an input layer, which is directly coupled to an input image and has the same number of neurons as the input image's pixels. Convolutional layers are the next set of layers, which display the results of convolving a series of filters with the input data and operate as a feature extractor. Each convolution layer's output is referred to as an activation map, which emphasizes the impact of applying a certain filter to the input. Finally, fully connected layers extract high-level abstractions [15].

As mentioned earlier, CNN-based algorithms have gone through teaks to improve their performance. A fully convolutional network (FCN) is one of the major improvements. In an FCN A fully convolutional layer was used in place of the last fully linked layer. This significant advancement enables the network to make dense pixel-by-pixel predictions [15].

Several studies such as [9] used a CNN model and an FCN model to conduct segmentation on several medical images datasets and in most cases, the FCN model had the better results. In a different study [17] an FCN model also shown superiority when conducting segmentation to detect diseases on images dataset of plants. In all studies, FCN models achieved a high accuracy of a minimum of 80 percent.

4.2.2 Traditional Machine Learning . Different studies [12], [3], [4] chose traditional machine learning techniques such as linear regression(LR), Random forest (RF), or support vector machine (SVM). In traditional machine learning, before choosing a model such as SVM or RF, a feature extraction method must be specified first. A filter bank containing a set of predefined filters is commonly used in feature extraction in machine learning. From the input data, such a filter bank can generate a variety of representations. After that, a feature selector should be specified to select the needed features over noisy or redundant features. The studies [12], [3], [4] achieved high results using traditional machine learning models. When there is a relatively small dataset, the traditional machine learning models outperform the deep-learning models.

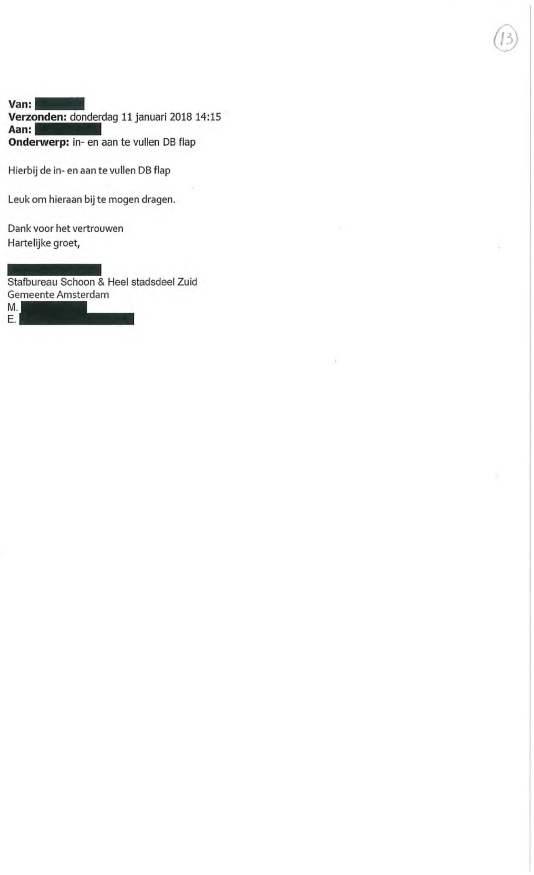
5 METHODOLOGY

This chapter is about methodology and discusses research strategies, data collection methods and data analysis methods.

5.1 Dataset

The dataset consists of 365 PDF files which contain a total of 32 thousand textual pages. The documents can be found and downloaded from the dutch government website webcovid19 (https://wobcovid19.rijksoverheid. The pdf documents contain information such as decisions from

Dutch ministries concerning corona regulations. The pdf documents mostly contain textual data but due to the privacy laws such as article 10 [16] that protects people personal data, non-linguistic elements are also present in the documents. Non-linguistic elements such as redacted regions that cover personal information.


(13)

Van: [redacted]
Verzonden: donderdag 11 januari 2018 14:15
Aan: [redacted]
Onderwerp: In- en aan te vullen DB flap

Hierbij de in- en aan te vullen DB flap

Leuk om hieraan bij te mogen dragen.

Dank voor het vertrouwen
Hartelijke groet,

[redacted]
Stafbureau Schoon & Heel stadsdeel Zuid
Gemeente Amsterdam
M. [redacted]
E. [redacted]

Figure 1: Example of redaction using quadrilateral shapes

5.2 Dataset preparation

To prepare the dataset, a few steps needed to be done. The first step was to write a script that would download the pdf documents from the government website. The next step would be to write a script that converts the downloaded pdf documents to images. Considering the relatively huge number of pdf documents, a common laptop would take up to two days to download the pdf documents and turn them into images.

Since not all pages pages contain redacted regions, an approach that only extracts pages that contain redaction was needed. The first approach that was tried to extract pages that only contained redaction was to write a script (in python) that detected certain shapes such as geometrical shapes since geometrical shapes were used in several of the downloaded PDF documents. This approach helped extract pages that contained redacted regions with sharp edges and geometrical shapes(such as quadrilateral shapes). As an example figure 1 is shown. However, there was an issue with this approach. This approach could not extract pages with different redaction styles from different sources (As an example check figure 2). As a result, the dataset missed several pages that contained different redaction styles. To solve this problem, another approach was chosen. For this approach, the PDF pages were reviewed by hand to look for all kinds of redaction styles and then label them using an annotation tool which was Label Studio. Label Studio is one of the most used tools for image segmentation, classification and detection purposes. This tool is used by well known companies such as Facebook and IBM [10].

Aan: [redacted]

Onderwerp: RE: Graag reactie uiterlijk 5 maart | Tijdelijk afvalpunt op Zeeburgereiland

Beste collega's,

Veel dank voor jullie hulp bij het invullen van de tabel. Ik heb nog een aantal extra alternatieven toegevoegd, met bij sommige een voorzet voor jullie reactie.
Zouden jullie daar nog snel een blik op willen werpen en zo mogelijk dinsdag 6 maart, uiterlijk woensdag 7 maart, willen reageren?

Ontzettend bedankt weer!

Met vriendelijke groet,

[redacted]

Gemeente Amsterdam

[redacted]

Werkdagen: ma, di, wo, do

Van: [redacted]
Verzonden: woensdag 28 februari 2018 16:06
Aan: [redacted]
CC: [redacted]
Onderwerp: Graag reactie uiterlijk 5 maart | Tijdelijk afvalpunt op Zeeburgereiland

Beste collega's,

Gisteren stond in de projectgroep Zeeburgereiland het verslag van het directieoverleg van 10 januari 2018 met de besluiten die zijn genomen ten aanzien van het tijdelijke afvalpunt en recyclecentrum op de bedrijvenstrook ter informatie geagendeerd. Het tijdelijke afvalpunt wordt een locatie met lage bakken omkleed met hoge hekken die afgewerkt worden in overleg met de buurt (met groen, doeken met kindertekeningen, ntb). Er is normaliter geen sprake van stankoverlast (geen huisvuil).

Met name vanuit [redacted] werden zorgen geuit over de aangewezen locatie nabij Haddock en relatief dichtbij de woningen van de Sportheldenbuurt. Er worden veel bezwaren vanuit de buurt verwacht. Uiteraard wordt de buurt voorafgaand op de planologische procedure gevraagd te reageren op het plan en zullen we deze participatie zorgvuldig voorbereiden samen met de RVE Grondstoffen i.o. met Q&A's. Ook staat Grondstoffen i.o. er voor open een beheeroverleg oid met bewoners in gevallen van overlast.

In de projectgroep Zeeburgereiland werden nog twee alternatieve locaties genoemd die mogelijk een optie kunnen zijn.
1. Haddock weg, afvalpunt op locatie Haddock
2. Afvalpunt naar busparkeerplaats P+R, bussen naar bedrijvenstrook (locatie OAT)

Ik zou jullie willen vragen de tabel in te vullen vanuit jullie expertise. Als je nog een andere locatie ziet op ZBE die past binnen de vereisten dan ook graag noemen en daarbij reactie vanuit jouw expertise.

4

Figure 2: Example of redaction that was done by hand.

5.3 Redaction detection model implementation

The first model was a 3-layer CNN model will be used on the images dataset, with the current image dataset a CNN model would have a high accuracy [14] . To optimize the output of the model, some regularization and an Adam are needed such as relu or sigmoid functions. For higher accuracy, the network should be trained independently on each type of image sample. If a new image shape is introduced, a new network can be established.

5.4 Experimental setup

5.4.1 CNN based model . Before building and training a CNN model that detects redaction in a dataset that consists of documents. The documents must be converted into images because models such as CNN work better on image features. After having the dataset ready. The dataset should be randomly split into a training, validation, and test set. To reduce the over-fitting, the hyper-parameter epochs can be set to 100 and the training will be stopped if the validation loss does not improve after 10 epochs. While building the CNN model, 3 layers must be added after that regularization functions and Adam must be added to optimize the output. The model was implemented using Python, TensorFlow, and Keras software.

5.4.2 Traditional machine learning . Traditional machine learning was built the same way as explained 4.2.2

5.4.3 Traditional machine learning with convolutional filters. Traditional machine learning was built the same way as explained 4.2.2 but as feature extractors, convolutional filters were used to see if it improves the results.

5.5 Evaluation

To measure the performance of the mode that handles segmentation problems, the panoptic quality(PQ) is used. The PQ is a combination of the IoU(Intersection over Union) and AP(Average Precision). The PQ consists of two steps. The first step is matching the predicted and the ground truth segment. The second step is mathematically calculating the PQ using the PQ formula shown in figure 3. [13]

$$PQ = \frac{\sum_{(p,g)\in TP} \mathrm{IoU}(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}$$

The Panoptic Quality

Figure 3: Panoptic quality formula

6 RESULTS

This section will discuss the results. All the models were subjected to the same dataset that was described in the Methodology section. Figure 4 shows a table of the averaged panoptic quality scores of the models. Judging by the table, the fully convolutional neural network model (FCN) achieved better results than the other models. In order to get an insight of the results, figure 5 shows a sample of the models' predictions compared to the ground truth images.

6.1 Results of the fully convolutional neural network model (FCN)

While building the fully convolutional model, it was needed to specify the network architecture which was achieved by specifying the backbone. The feature extractor network employed within the CNN architecture is referred to as the backbone. The network's input is encoded into a feature representation using this feature extractor [8]. The Resnet34 backbone achieved better results than the rest of the backbones. In addition the Adam optimiser and the jaccard loss function was implemented to eventually get the the current results.

6.2 Results of traditional machine learning

As explained in section 4.2, before specifying a model such as Rain forest to segment images, a feature extraction function had to be specified first by building a filter bank which consists of several

feature extracting filters. The authors of [12], [3], [4] implemented the Gabor, Gaussian and Canny edge extraction filters in Rain forest and Support vector machine models to conduct segmentation on medical images. Those filters and traditional machine learning models did not achieve the desired goal to segment the redacted regions on the dataset from the methodology section. After that, more filters were implemented to test if traditional machine learning would segment the redacted regions. At the end this approach did not achieve the desired results as shown in figure 4 and figure 5.

6.3 Results of traditional machine learning with convolutional filters

Similar to the approach in the above section, same machine learning models were implemented but instead of using the same filter extraction filters, we added convolutional filters as feature extractors. The models in this approach did not achieve the desired goal to segment the redaction regions as shown in figure 4 and figure 5.

Models	Panoptic quality	IoU score
Model 1 (Fully convolutional neural network model (FCN) - deep-learning)	0.70	0.82
Model 2 (Rain forest (RF) - Traditional machine learning)	0.0	0.0
Model 2 (Support vector machine (SVM)- Traditional machine learning)	0.0	0.0
Model 3 (Rain forest (RF) - Traditional machine learning with convolutional filters)	0.0	0.0
Model 3 (Support vector machine (SVM)- Traditional machine learning with convolutional filters)	0.0	0.0

Figure 4: Results of models

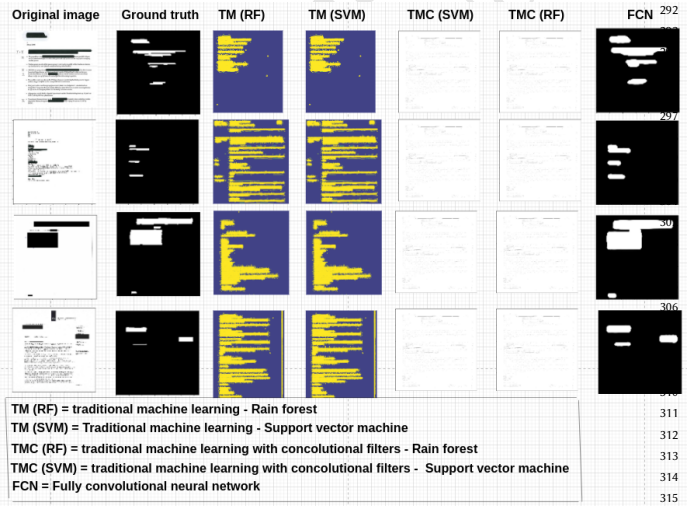


Figure 5: visualisation of results

7 DISCUSSION

7.1 Results of the models

In this research, an attempt was made to detect redacted regions in documents. Different approaches were tried using the state of the art deep learning and machine learning models. Out of all the approaches, the fully convolutional model (deep-learning) worked best for the given dataset. The other models preformed poorly as shown in the results section.

7.2 Limitation

There are few limitations in this study. The first one is not having a small dataset. Despite having thousands of pages thanks to the WOO documents, only a relatively small amount of pages contained redacted regions. This is because redaction was typically only needed when a page contained private details such as names or phone numbers which usually occurred only a few times in each document. This limitation can be solved by augmenting more images. Another limitation is that in the related work, the machine learning performed well based on the features their datasets had. All their dataset were non-textual (such as medical images), thus they had more features that the dataset in this research misses.

8 CONCLUSION AND FUTURE WORK

In this research, an attempt were made to segment and detect redacted regions in WOO documents. In the related work, state of the art CNN and traditional machine learning based models were used to segment images and detect certain objects. for the given dataset, the FCN based model worked the best. For future work, more professionally labelled data is needed since deep-learning is the only was to tackle this problem because traditional machine learning did not work on this kind of dataset.

REFERENCES

[1] J. Van Iden A. Kleinhout, N. van den Biggelaar. NETHERLANDS INTRODUCES NEW ACCESS TO PUBLIC INFORMATION REGIME. 2022.
[2] Adobe. Why redact or remove sensitive content? 2021.
[3] Chee-Kong Chui andSim-Heng Ong Agus Pratondo. Integrating machine learning with region-based active contour models in medical image segmentation . 2016.
[4] Yusuf Artan. Interactive Image Segmentation using Machine Learning Techniques . 2011.
[5] business.gov.nl. business.gov.nl. 2022.
[6] Kam Woods Christopher (Cal) Lee. Automated Redaction of Private and Personal Data in Collections. 2012.
[7] Dan Gebow David A. Clunie. Block selective redaction for minimizing loss during de-identification of burned in text in irreversibly compressed JPEG medical images. 2015.
[8] Rongyu Zhang et al. Comparison of Backbones for Semantic Segmentation Network . 2020.
[9] Jimin Liang-Xueli Chen Getao Du, Xu Cao and Yonghua Zhan. Medical Image Segmentation based on U-Net: A Review . 2020.
[10] Heartex. labelstud. 2022.
[11] Jia W. He X. et al Hesamian, M.H. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. 2019.
[12] Varun Vasudevan Charles Huang Hongyi Ren Ruoxiu Xiao Xiao Jia Lei Xing Hyunseok Seo, Masoud Badiei Khuzani. Machine learning techniques for biomedical image segmentation: An overview of technical aspects and introduction to state-of-art applications . 2020.
[13] Harshit Kumar. Introduction to Panoptic Segmentation: A Tutorial. 2019.
[14] Prafful Mishra. Why are Convolutional Neural Networks good for image classification? 2019.
[15] 2 · Wenjing Jia1 · Xiangjian He1 · Paul Kennedy3 Mohammad Hesam Hesamian1. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges . 2019.

[16] Ministry of the Interior and Kingdom Relations. De nederlandse Grodwet. 2022.
[17] Yash Paul Singh Berwal Parul sharma and Wiqas Ghai. Performance analysis of deep learning CNN models for disease detection in plants using image segmentation. 2020.
[18] ISHITA DAS SWARNENDU GHOSH, NIBARAN DAS and UJJWAL MAULIK. Understanding Deep Learning Techniques for Image Segmentation . 2019.
[19] Bernt Schiele Tribhuvanesh Orekondy, Mario Fritz. Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images. 2015.
[20] Yan Wang Xiang Bai Zhijiang Zhang Kurt Keutzer Wei Shen, Xinggang Wang. DeepContour: A Deep Convolutional Feature Learned by Positive-sharing Loss for Contour Detection. 2015.
[21] Deliang JIANG Xiaohu YANG. Converting PDF to HTML approach based on Text Detection . 2009.
[22] Ke Gao1 Sheng Tang1 Yongdong Zhang1 Yang Cao1, 2. A Representative Local Region Detector Based On Color-Contrast-MSER . 2014.