# 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, 5 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. Deeplearning 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 literature, several approaches have been used to detect or extract features and information from PDFs such as region detection, image segmentation, colors detection and shape detection. For example, [14] presents a representative local region detector based on color-contrastpixel ranking and [13] proposes extracting more information and features from PDFs by converting PDFs to HTML form. Furthermore, different research papers proposed different ways to extract information from PDFs after turning them into images. Such as using a Convolutional Neural Network model to extract image features. What most of the research papers have in common is that they used the extracted image features such as shape or color to detect certain objects or to classify types of documents. What was insufficient from the literature is conducting features extraction and / or segmentation on textual documents to detect and classify redaction. In this paper, an attempt was made to successfully detect whitewashed text in pdfs documents by applying deeplearning and machine learningtechniques. 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) [7] when detecting and calculating the percentage of redacted regions in FoIA pages? **SQ1:** How do baseline models based on [12] preform on FoIA 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 [11] and Convolutional filters with machine learning? **SQ 3:**How do these models perform when applied to FoIA 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 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 [3]. 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 [1]. Now more on the redaction process. When the information that needs to be redacted is identified. There are two approaches mentioned by [4]. 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, [11] 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 [11] labeled the ground truth image dataset by hand to ensure accuracy.

## 4.2 Contour and Edge Detection

First, let's give a definition to edge detection. The method of edge detection seeks to capture the important properties of objects in an image. Discontinuities in the photometrical, geometrical, and physical qualities of objects are examples of these properties. Variants in the grey level image result from this information; the most common variations are discontinuities (step edges), local extrema (lines edges), and 2D features generated when at least two edges meet (jounctions) [8].

Since that several redacted regions in documents are presented in geometrical like shapes. Several papers that aimed to detect certain shapes or edges are discussed. In [5] a method for accurately detecting two-dimensional (2-D) forms has been suggested. The shape boundary's cross section is described as a step function. First, a one-dimensional (1-D) optimal step edge operator was developed that reduces both noise power and mean squared error between the input and filter output. This approach of detecting a form is a natural extension of the problem of pixel-level edge detection to the problem of global contour identification. This simple filtering approach also serves as a tool for analyzing edge-based shape identification in a systematic manner. For detecting a certain shape, contour detection techniques should be investigated. It is pointed out in [2] that contour detection is fundamental in image segmentation and object detection and classification. Instead of old-fashioned simple filtering, the authors chose to build a detector by comprising parallel k-means, convolution, and skeletonization routines in addition to the local cues and eigensolver routines. Combining these contributions helped reduce the run-time.

Improving the accuracy for contour detection by using deep learning instead of old fashioned simple filtering also worked in [12] where a CNN model was trained on a multi-class classification task, namely to classify an image patch to which shape class or the negative class. The goal of training a standard CNN is to maximize the probability of the correct class, which is achieved by minimizing the softmax loss.

# 5 METHODOLOGY

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

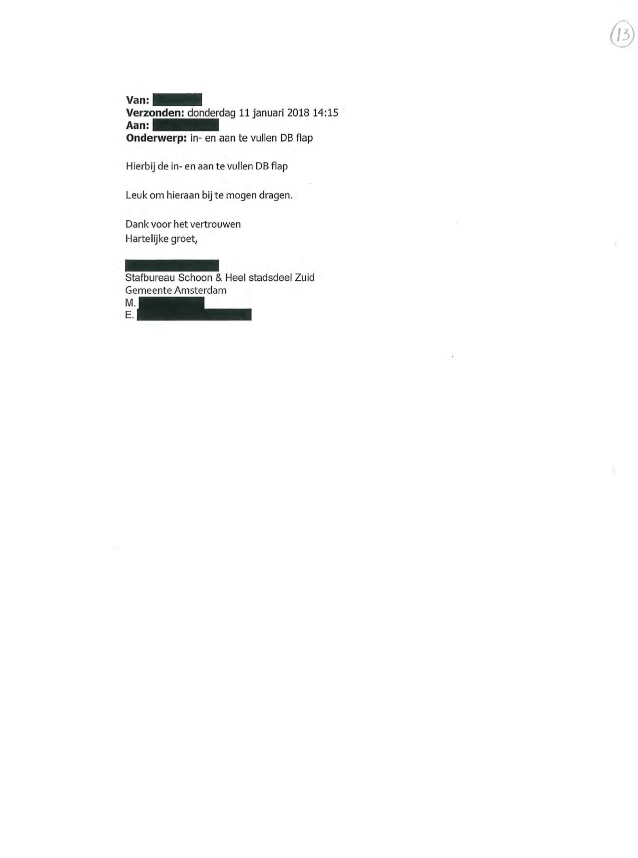
## 5.1 Dataset

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 [10] 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.

## 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 [6].



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

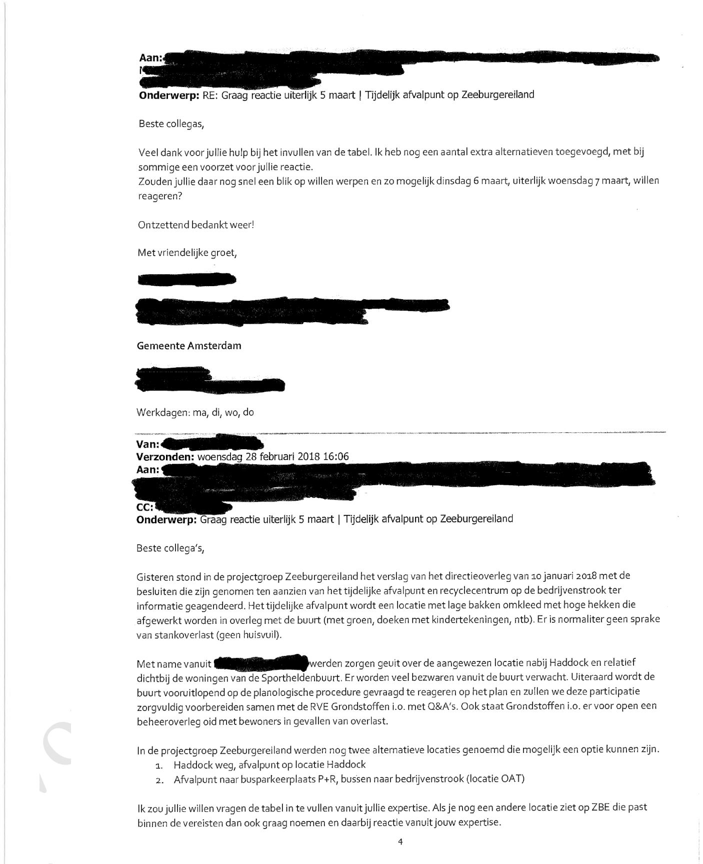
Hierbij de in- en aan te vullen DB flap

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**



Aan: [redacted]
Onderwerp: RE: Graag reactie uiterlijk 5 maart | Tijdelijk afvalpunt op Zeeburgereiland

Beste collegas,

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 vooruitlopend 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 [9] . 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

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.

## 5.5 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. [7]

# 6 RESULTS

This section will discuss the results.

| | Accuracy | Segmentation Metrics |
|---|---|---|
| **Model 1** | 3 % | |
| **Model 2** | 93 % | |
| **Model 3** | 94 % | |

**Figure 3: Results of models**

# 7 DISCUSSION

# 8 CONCLUSION

# REFERENCES

[1] Adobe. Why redact or remove sensitive content? 2021.
[2] Narayanan Sundaram Yunsup Lee Mark Murphy Kurt Keutzer Bryan Catanzaro, Bor-Yiing Su. Efficient, High-Quality Image Contour Detection. 2009.
[3] Kam Woods Christopher (Cal) Lee. Automated Redaction of Private and Personal Data in Collections. 2012.
[4] 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.
[5] Azriel Rosenfeld Hankyu Moon, Rama Chellappa. Optimal Edge-Based Shape Detection. 2002.
[6] Heartex. labelstud. 2022.
[7] Harshit Kumar. Introduction to Panoptic Segmentation: A Tutorial. 2019.
[8] Honglian Ma Lei Zhai, Shouping Dong. Recent Methods and Applications on Image Edge Detection. 2008.
[9] Prafful Mishra. Why are Convolutional Neural Networks good for image classification? 2019.
[10] Ministry of the Interior and Kingdom Relations. De nederlandse Grodwet. 2022.
[11] Bernt Schiele Tribhuvanesh Orekondy, Mario Fritz. Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images. 2015.
[12] 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.
[13] Deliang JIANG Xiaohu YANG. Converting PDF to HTML approach based on Text Detection . 2009.
[14] Ke Gao1 Sheng Tang1 Yongdong Zhang1 Yang Cao1, 2. A Representative Local Region Detector Based On Color-Contrast-MSER . 2014.