University of Glasgow | School of Computing Science

Honours Individual Project Dissertation

Why is This Sensitive? Visualising Important Sensitivity Classification Features

Guillaume de Susanne d'Epinay

March 9, 2020

Abstract

Every abstract follows a similar pattern. Motivate; set aims; describe work; explain results.

“XYZ is bad. This project investigated ABC to determine if it was better. ABC used XXX and YYY to implement ZZZ. This is particularly interesting as XXX and YYY have never been used together. It was found that ABC was 20% better than XYZ, though it caused rabies in half of subjects.”

Education Use Consent

I hereby grant my permission for this project to be stored, distributed and shown to other University of Glasgow students and staff for educational purposes. **Please note that you are under no obligation to sign this declaration, but doing so would help future students.**

Signature: Guillaume de Susanne d'Epinay Date: March 9, 2020

Contents

1 Introduction

1.1 The need for Technology Assisted Sensitivity Review

Motivations why senstivity review why technological problem

1.2 Objectives

1.3 overview

bulle tpoi t of chapter ocntent

2 Background

need for explanations in ML

CITE SOME WORKS

2.1 Literature Review

2.1.1 Technology Assisted Review

McDonald et al. (2014) A first implementation of Machine Learning classifier for sensitivity Review

McDonald et al. (2015) Improving classification with Part of Speech tagging

McDonald et al. (2017) Improving classification with word embeddings

McDonald et al. (2018) Active Learning ("loopback learning") implemeting reviewer feedback in Sensitivity prediction

McDonald et al. (2019) User study with Machine Learning classification

CITE SOME GENERAL ML EXPLANATIONS PAPERS

2.1.2 Document visualization techniques

and elaborate on why didnt choose them "people have done this"

2.2 Related products

Quite a few official government guidelines on the redaction of sensitive documents mention Adobe Acrobat as a go to tool for redacting text documents The National Archives (2016). Sometimes, the guidelines *are* Adobe's guidelines for redaction Scottish Government (2019).

2.2.1 PDF editors

Adobe Acrobat is a well known PDF toolkit that notably enables sensitivity redaction of documents. The principle is common: select text, click redact and browse a nested context menu to select an exemption (Figure 2.1).

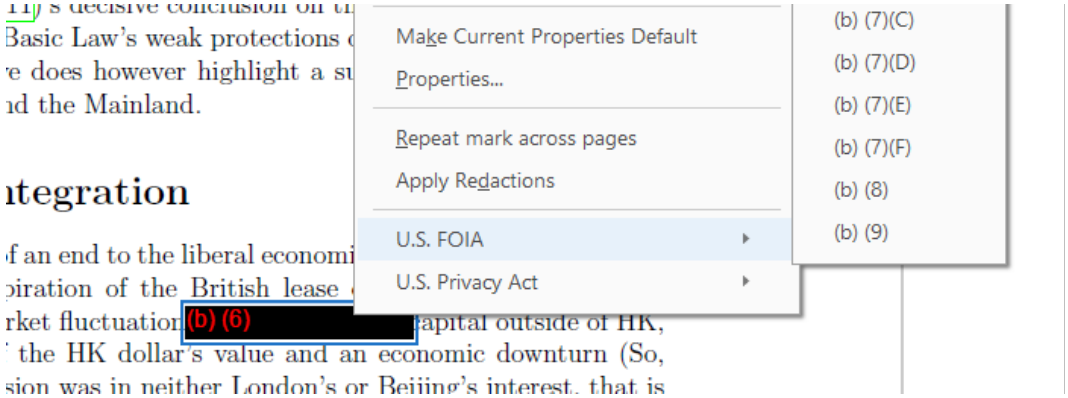


Figure 2.1: *Adobe Acrobat redaction workflow*

There are a variety of PDF editors, when they implement document redaction, they closely resemble this workflow. For instance, Foxit PhantomPDF is another well known PDF editor that implements a strikingly similar context menu (Appendix A.1) also allows for document wide redaction of a text selection (Appendix A.2), something which Adobe implement quite poorly.

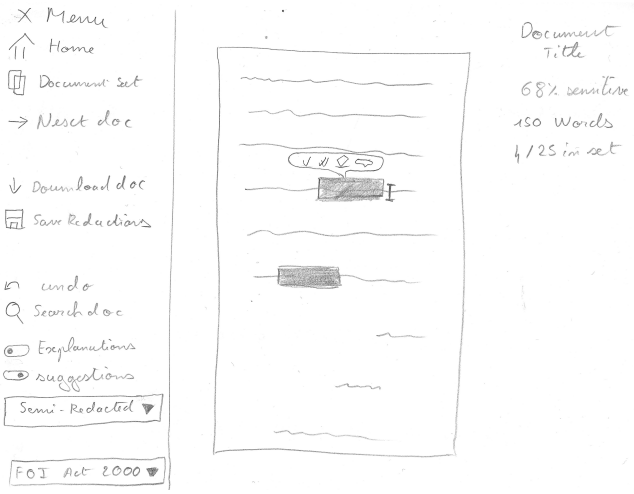
3 Requirements

3.1 Scope

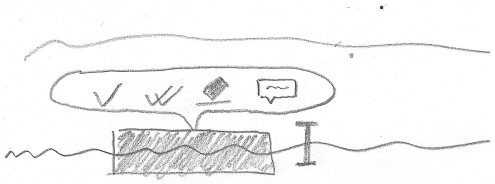
3.2 (non)Functional definition

user study

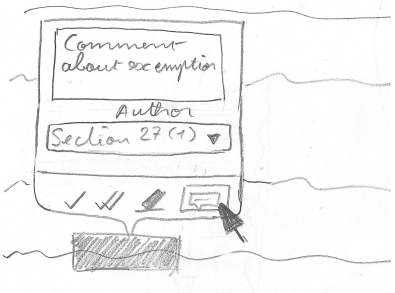
3.3 Wireframes



(a) Document View Wireframe



(b) minimized tooltip wireframe



(c) Expanded tooltip wireframe

Figure 3.1: *Selection of some of the Wireframes drawn up during the design process*

4 Design

link back to requirements
choice of technologies

4.1 Application Structure and Technology stack

4.2 Machine Learning Model

4.3 API

4.4 Frontend

5 Implementation

show coponennt code and how it gets data from backend

show lime explanations generation

key "features" code overview

- Hands on SkLearn framework, experimented with different text pre-processing techniques and classifiers to build a Pipeline to classify textual documents into either sensitive or not sensitive categories
- overview of Machine Learning model explanation techniques, successfully implemented the Lime explainer, tried to use the SHAP explainer, but was unsuccessful
- Explore text document storage, looked at Terrier, ElasticSearch to finally settle on MongoDB
- Packaged SkLearn classifier into Python Flask API with OpenAPI specification
- created Javascript API Client from code generated from OpenAPI specification
- hands on ReactJS Javascript frontend framework, implemented a document browsing, viewing and upload frontend
- added text redaction feature to mark sections as sensitive with a label explaining why it is deemed sensitive
- implemented Lime Model explanations in Python backend, exposed them to the Flask API and created a frontend "in text" visualization of these sensitivities
- Research Document visualization techniques (literature review)
- created a populator script to load and classify all documents
- graph visualization of feature contribution to the classification and other UI tweaks

6 Evaluation

We conducted a user evaluation in order to evaluate the effectiveness of our application in helping reviewers conduct a sensitivity review. We sought to answer a handeful of research questions.

- Firstly, does our visualization of predicted document sensitivity and explanation features help reviewers conduct a sensitivity review faster?
- Secondly, does it improve the accuracy of the sensitivity review process.
- Lastly, does it improve reviewers' confidence in the completeness of their sensitivity review

6.1 Preparations and Experimental Setup

The collection is a set of 3801 Government documents relating to International Activities. Our collection's ground truth for sensitivity was established with respect to sections 40 (International Relations) and 27 (Personal Information) of the Freedom of Information Act (FOIA).

An important consideration in the decisions that follow is that we did not have access to professional document reviewers, as such the evaluation was conducted with students in the role of reviewers. Out test reviewers were asked to identify sensitivities according to section 27 of the FOIA. Since our reviewers are not experts, thus finding Personal Information in a document is easier than identifying sensitivities harming International Relations (Section 40). Due to time constraints we only selected short documents to show our reviewers, that is, documents under 2000 characters.

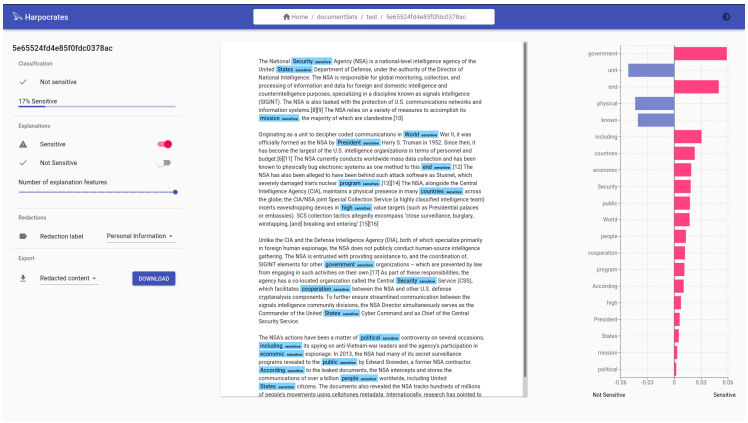
We selected both sensitive and non-sensitive documents in order to represent every category of the confusion matrix for the classifier with the following counts:

	Actually non-sensitive	Actually Sensitive
Predicted non-sensitive	3	1
Predicted Sensitive	1	1

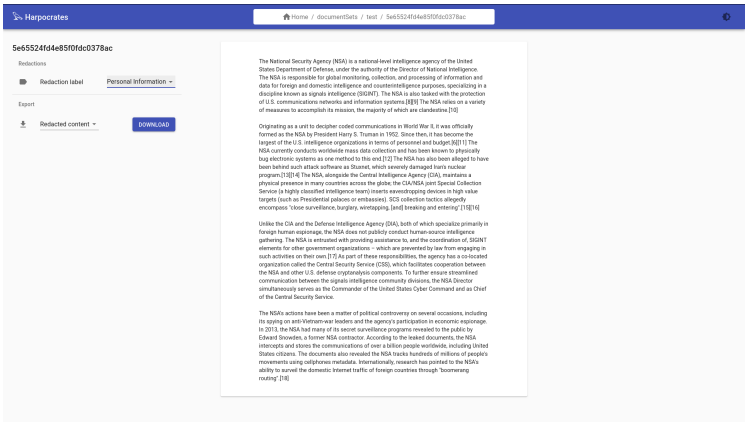
Table 6.1: *Document selection in the classifier's confusion matrix*

Multiple interfaces were required to evaluate the effectiveness of the application. One interface *test mode 1* (Figure 6.1a) is a simplified version of the final interface containing the classification prediction and the accompanying explanations. The other, *test mode 2* (Figure 6.1b) is a stripped down version that only displays the document, and the manual redaction tools (document title and sensitivity type selection).

Our reviewers reviewed documents with both interfaces, the first interface displayed was chosen at random.



(a) Test mode 1



(b) Test mode 2

Figure 6.1: *The two test modes used for evaluation*

Lastly, after the review of a batch of documents with each interface, a questionnaire was handed out to the testers

6.2 Evaluation

As mentioned, each user will be asked to review one collection of 6 documents for personal information sensitivities (FOIA Section 27) per user interface

The independent variables will a set of two user interfaces: with and without the predicted classification and explanations (more details below) which all users will both use. We will measure multiple dependent variables:

- Firstly, we will measure the time to review an entire collection with each interface.
- Secondly we will record the accuracy of each reviewer on each interface for all documents.
- Lastly, we will evaluate the confidence of the reviewers after using each interface with a Likert scale in the questionnaires.

Research questions dependent/indepdent variables UI variations experimental setup document, subject seleciton

results

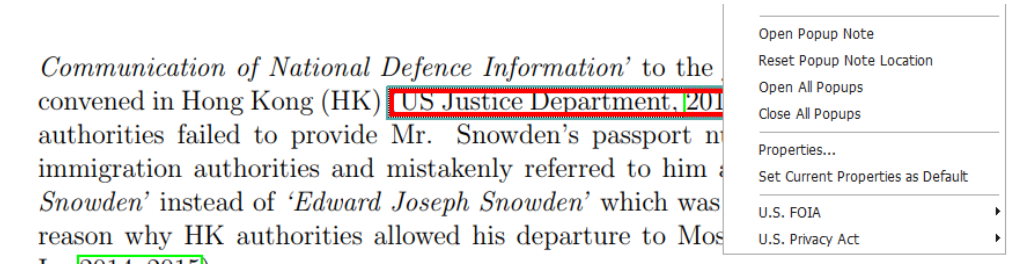
link back to requirements how have they been met unit testing

7 Conclusion

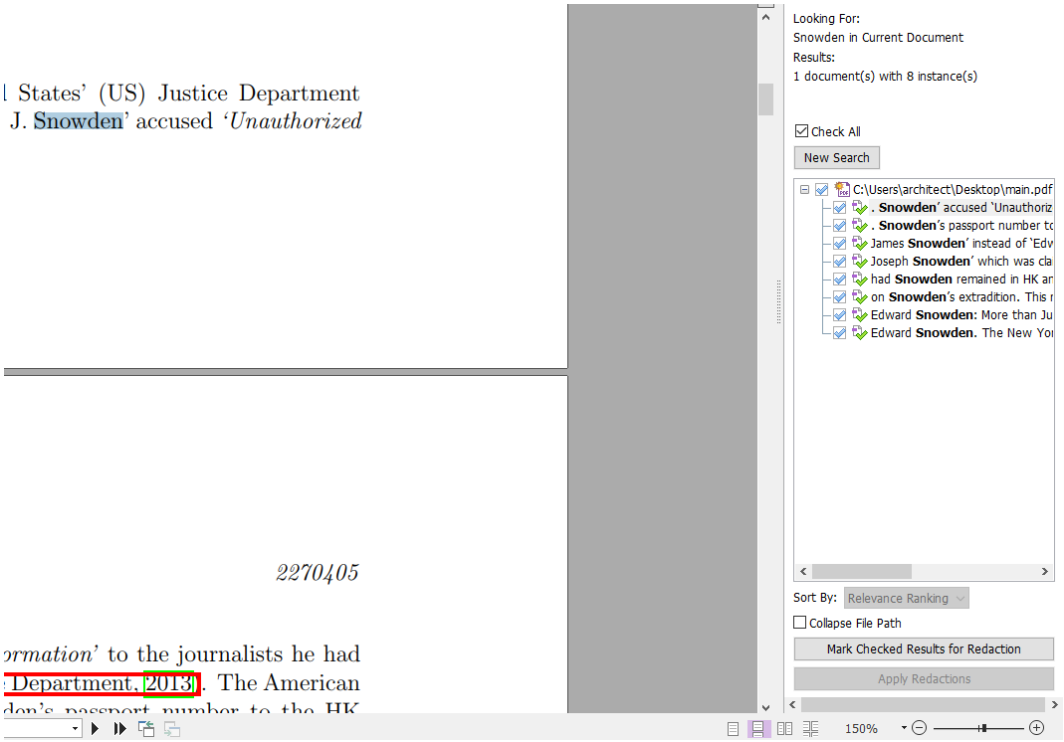
requirements I've met
reflections
future work

A PDF Editors

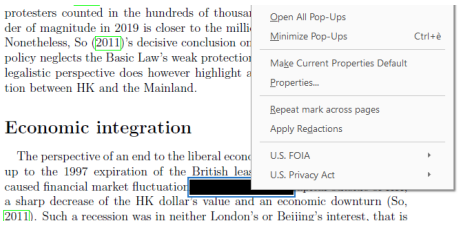
A.1 PhantomPDF context menu



A.2 PhantomPDF redact all



A.3 Adobe Acrobat redact all



protesters counted in the hundreds of thousa
der of magnitude in 2019 is closer to the milli
Nonetheless, So (2011)'s decisive conclusion o
policy neglects the Basic Law's weak protectio
legalistic perspective does however highlight a
tion between HK and the Mainland.

Economic integration

The perspective of an end to the liberal econ
up to the 1997 expiration of the British lea
caused financial market fluctuatio
a sharp decrease of the HK dollar's value and an economic downturn (So, 2011). Such a recession was in neither London's or Beijing's interest, that is

latter two respectively competing with HK's services and financial sectors (Chan, 2019; X. Chen, 2018).

Despite some economic downturns, it is not the growth rate of HK that has declined, rather, the share of its contribution to the Chinese Gross Domestic Product (GDP) has fallen from 13% in 1992 to 2% in 2012, is a better illustrator of the new power paradigm in the Mainland's relationship to HK (Yeung & Huang, 2015, pp. 193–194). It is in part this shift that has conferred Beijing more confidence to assert its 'One China' agenda of sovereignty over HK (So, 2011, p. 113).

Democracy with Chinese characteristics

This first point of the Joint Declaration affirms the resumption of the PRC's sovereignty over HK on the 1st of July 1997 (People's Republic of

7 Bibliography

G. McDonald, C. Macdonald, I. Ounis, and T. Gollins. Towards a Classifier for Digital Sensitivity Review. In M. de Rijke, T. Kenter, A. P. de Vries, C. Zhai, F. de Jong, K. Radinsky, and K. Hofmann, editors, *Advances in Information Retrieval*, volume 8416, pages 500–506. Springer International Publishing, Cham, 2014. ISBN 978-3-319-06027-9 978-3-319-06028-6. doi: 10.1007/978-3-319-06028-6_48.

G. McDonald, C. Macdonald, and I. Ounis. Using Part-of-Speech N-grams for Sensitive-Text Classification. In *Proceedings of the 2015 International Conference on Theory of Information Retrieval - ICTIR '15*, pages 381–384, Northampton, Massachusetts, USA, 2015. ACM Press. ISBN 978-1-4503-3833-2. doi: 10.1145/2808194.2809496.

G. McDonald, C. Macdonald, and I. Ounis. Enhancing Sensitivity Classification with Semantic Features Using Word Embeddings. In J. M. Jose, C. Hauff, I. S. Altıngovde, D. Song, D. Albakour, S. Watt, and J. Tait, editors, *Advances in Information Retrieval*, volume 10193, pages 450–463. Springer International Publishing, Cham, 2017. ISBN 978-3-319-56607-8 978-3-319-56608-5. doi: 10.1007/978-3-319-56608-5_35.

G. McDonald, C. Macdonald, and I. Ounis. Active Learning Strategies for Technology Assisted Sensitivity Review. In G. Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury, editors, *Advances in Information Retrieval*, volume 10772, pages 439–453. Springer International Publishing, Cham, 2018. ISBN 978-3-319-76940-0 978-3-319-76941-7. doi: 10.1007/978-3-319-76941-7_33.

G. McDonald, C. Macdonald, and I. Ounis. How Sensitivity Classification Effectiveness Impacts Reviewers in Technology-Assisted Sensitivity Review. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval - CHIIR '19*, pages 337–341, Glasgow, Scotland UK, 2019. ACM Press. ISBN 978-1-4503-6025-8. doi: 10.1145/3295750.3298962.

Scottish Government. Redacting Information. Technical Report foi-19-00783, Apr. 2019.

The National Archives. Redaction toolkit for paper and electronic documents: Editing exempt information from paper and electronicdocuments prior to release. Technical report, The National Archives, Apr. 2016.