Visualising Important Sensitivity Classification Features

Mohib Akoum - 2431135a

December 17, 2021

1 Status report

1.1 Proposal

1.1.1 Motivation

Freedom of information laws legislate that public entities must release documents to the public. This includes government departments, such as the Ministry of Defence. However, because some of these documents contain sensitive data, they need to be filtered to show only the non-sensitive documents to the public. Therefore, an artificial intelligence agent that helps classify documents' sensitivity is crucial for speeding up the release of documents.

1.1.2 Aims

The project aims to develop a website that automatically displays the classifier's prediction about a document's sensitivity. However, due to the sensitivity of this task and the chance of false predictions, the website also displays explanations about why the classifier predicted a document as (non-)sensitive. The explanations will reveal which words impact the classification and whether that word increases or decreases the perceived sensitivity of the document. Moreover, The website should produce useful visualisations that explain how the classifier makes its predictions and the highest impact terms across the corpus.

1.2 Progress

- Language and web framework chosen: I used Flask with Python programming language.
- Classifier trained on corpus and capable of predicting a document's sensitivity.
- Lime explanations used to highlight the terms affect the classifier's decision making.
- Fixed imbalanced predictions caused by 85% of data set being sensitive documents.
- Used Cross-validation for evaluating the model.
- Used Cross-validation to allow the user to view all documents alongside the classifier's predictions.
- Added Shap to give additional visualisation alongside Lime.
- Full test coverage of all lines at the end to end system using 34 tests.

1.3 Problems and risks

1.3.1 Problems

Problems previously experienced:

- Implementing Lime with a saved model instead of a newly trained one.
- Classifier predicting all documents as non-sensitive due to class imbalance.
- Failure implementing BERT on full data set due to size of each document and small available ram.
- Unsupported and outdated libraries when attempting to implement some deep learning models.

1.3.2 Risks

Future problems:

- The addition of new features and explanations could result in long wait times for the user. Mitigation: optimise, and save previously created explanations.
- Classifier sometimes makes false predictions. Mitigation: Improve classifier by potentially tuning hyperparameters or using a different algorithm.

1.4 Plan

- Week 1-2: Implement additional explanations for classifier decision making on the document level as well as the corpus level
- Week 3-4: Attempt to get BERT to work on data set through the use of a cluster (higher ram than available on my personal machine). Furthermore, implement multiple models and compare their predictions.
- Week 5: Fully test end to end product and ensure robustness.
- Week 6-7: Run evaluation experiments to better understand the usefulness of the product created.
- Week 6-10: Write a draft for the dissertation and submit it to the supervisor for review.

1.5 Ethics and data

Since the project is intended to help differentiate the sensitivity of documents, evaluations can help measure the classifier's effectiveness and explanations. However, only a specialised few should be targeted for the evaluation due to the project's sensitivity.

In the first week of the second semester, I will seek guidance from my supervisor for what ethical approval will be necessary to complete the evaluation.