Diagram

Description automatically generated**Method One – Sudo Document Training**

A corpus of documents with existing truth labels are used as training data. Each document is used to create a sudo document containing a concatenation of the abstracts of each entity instance within the document. The features are extracted from each of the training sudo documents and passed into a learning algorithm. The learning algorithm is used by a classifier to generate predictions on standalone-unlabelled entity abstracts. The output of the classifier is the probability prediction of whether the abstract input is sensitive or not.

**Method Two – Lone Abstract Training**

Diagram

Description automatically generatedA set of distinct entities and their abstracts are retrieved from a labelled corpus of documents. Entity abstracts are assigned the truth value of the document they appear in. However, an entity can appear in both sensitive and non-sensitive documents. For example, if an entity appears in a document labelled as sensitive – then this instance of an entity abstract is labelled sensitive. If it appears in a non-sensitive document, then a separate instance of the abstract is assigned a truth label of non-sensitive. Features are extracted from each entity abstract, with a learning algorithm utilising k-fold cross-validation to make sensitivity predictions for each entity.

**One-Hot vs TF-IDF Feature Extraction**

When extracting features from entity abstracts or sudo documents I was most comfortable with one-hot encoding however I recognised that TF-IDF was a viable option in this scenario. One-hot encoding involves representing a document as a vector of binary values of whether each word in the vocabulary appears in a given document or not. TF-IDF encoding utilises the frequency of documents of which a given word does not appear in (IDF) as well as the frequency of each word in a given document (DF) to produce a sparse vector representation of each word in a given document. One-hot encoding does not take into account frequency of words due to the binary nature of the vector representation, whereas TF-IDF encoding presents a far more effective way of representing the frequency and relevancy of words in a given document. Given that capturing the relevance of certain words in entity abstracts was crucial, I felt TF-IDF encoding was more suitable in this scenario.

**Down Sampling Training Data**

In the process of experimenting with both classification methods, I noticed that the results were appearing heavily skewed towards predicting unsensitive labels. Upon inspecting class labels within the training data, unsensitive documents were far more frequent than documents labelled sensitive. Assuming this was the cause of the skewed predictions, I applied down sampling to the training data. Down sampling is the process of randomly sampling observations from the more frequent class label such that the number of observations used in the training data is equal for each class label. By randomly sampling an equal number of entities assigned an unsensitive label as the frequency of entities that were labelled sensitive, the training data resulted in far more intuitive predictions from the trained model.

**K-Fold Cross Validation**

**Feature Set Combinations & Abbreviations**

|  |  |  |
| --- | --- | --- |
| **Feature Set** | **Sudo Document** | **Individual Abstract** |
| Baseline | SD | IA |
| w/ Down Sampling | SD-DS | IA-DS |
| w/ Cross Validation | SD-CV | IA-CV |
| w/ Down Sampling & Cross Validation | SD-DS&CV | IA-DS&CV |

**Results**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Configuration** | **Precision** | **Recall** | **F1** | **F2** | **BAC** |
| SD | 0.6145 | 0.0400 | 0.0751 | 0.0492 | 0.5170 |
| SD-DS | 0.3785 | 0.5960 | 0.4630 | 0.5346 | 0.5938 |
| SD-CV | 0.0650 | 0.0404 | 0.0498 | 0.0437 | 0.4761 |
| SD-DS&CV | 0.5176 | 0.6545 | 0.5781 | 0.6216 | 0.5222 |
| IA | 0.7170 | 0.0099 | 0.0196 | 0.0124 | 0.5045 |
| IA-DS | 0.3332 | 0.5642 | 0.4189 | 0.4955 | 0.6456 |
| IA-CV | 0.1482 | 0.0534 | 0.0785 | 0.0612 | 0.4991 |
| IA-DS&CV | **0.5392** | **0.6972** | **0.6082** | **0.6587** | **0.5508** |

**Conclusion**