Post Block Assignment 3

Applied Machine Learning
Jeandre de Bruyn
19768206

3. Case study Document classification

Introduction:
The process of moderating academic papers submitted to Arxiv is a laborious process when it comes to making sure that the papers claimed subject area matches the correct subject area. In order to save time and man power a classification model will be developed which will categorize an article as belonging to the section Information theory or belonging to a different section.

Objectives and validation strategy:

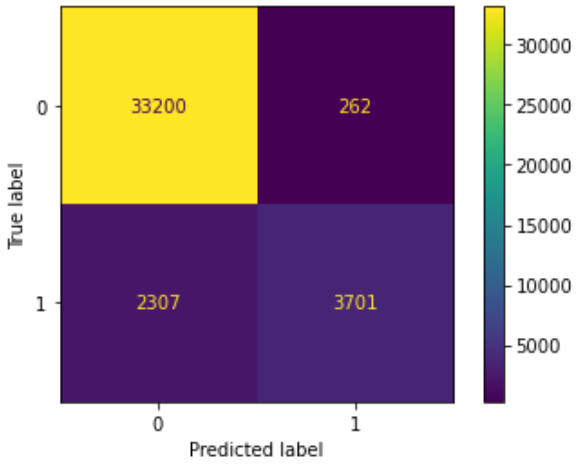
In order to achieve good performance in the classification model feature extraction will have to be performed. Therefore, one of the first objectives is to perform feature extraction that leads to high predicative power for the classifier. The raw data that is available to work with are ID, Date, Title, Abstract and Subject Area. The subject area is our target label and so our information available for feature extraction is limited to the other given information. Second objective after feature extraction is the training of a model and its evaluation. The evaluation of the model will be done using Accuracy, Precision and Recall (for the positive case and the negative case). The metric that will be the most important to focus on will be positive class recall as it is the minority class, and we want to maximize correctly identified Information Theory articles. The model will be evaluated on its generalization ability using a train/test split of 70/30.

Data Preprocessing:
Since the date and ID features provide no useful information as to the subject area of the article these two features are discarded and the only two features left to use are title and abstract. These two features are joined per article to create one “text” feature per article. Feature extraction is done using a bag of words implementation from sklearn “CountVectorizer”. The bag of words is fitted to the training data set and then both the training and test sets are transformed into their respective vector representations.

Initial model and evaluation:

Table 1: Initial model performance metrics

Accuracy	0.8041
Precision	0.9339
Positive recall	0.6160
Negative recall	0.9922



The initial model performs very well in terms of negative recall with 99.2% and only 262 articles misclassified as positive instead of negative. However this is overshadowed by rather poor performance in positive case identification with only 61.6% being correctly identified. Overall accuracy is good at 80% but due to the interest being in the positive case identification it would be preferred to have a higher positive recall. These results are to be expected considering the severe class imbalance with the positive case only representing 15.2 % of the cases.

In order to improve the performance of the model the feature extraction will have to be improved. In the current iteration the entire list of words available in the training set is considered as the vocabulary, this makes it both hard to train as well as having a lot of unnecessary information being fed as training data to the random forest classifier. There are a lot of unique terms and words present in the titles and abstracts which just act as noise and overfit the model. Since Random forest classifier is unpruned the solution to this problem is to be more selective when it comes to the creation of the bag of words model. The following changes should be made to the bag of words model to improve performance:

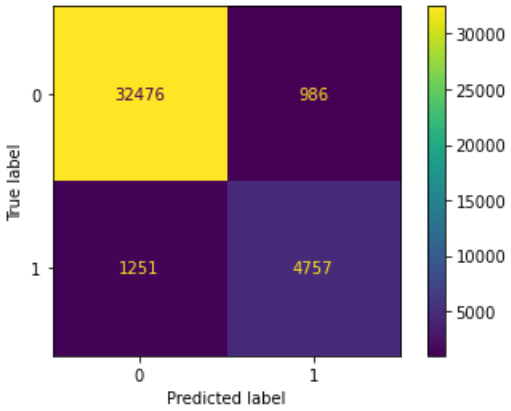
- Require a minimum word frequency to be in the bag of words, this removes any unique occurrences that add noise to the data.
- Require that a word only be present in less than a certain percent of all articles, this eliminates any general words that are commonly used and don't help to differentiate between the subject areas.
- Use ngrams in order to provide context per sentence. Helps provide differentiation between "this is good" and "this is not good".

- Remove stopwords which provide no differentiating information between subject areas.
- Have a maximum amount of features/words that are in the vocabulary. This initial model has over 100 000 words, this promotes overfitting.

Optimal Model:
The final model was obtained by following the above-mentioned possible improvements to the feature extraction method used. Maximum number of words is set to 30 000, stop words are removed, bigrams are used rather than individual words, a word has to occur at least twice to be relevant as well as be present in less than 80% of all articles.
The evaluation of the final model is as follows:

Table 2: Final model evaluation metrics

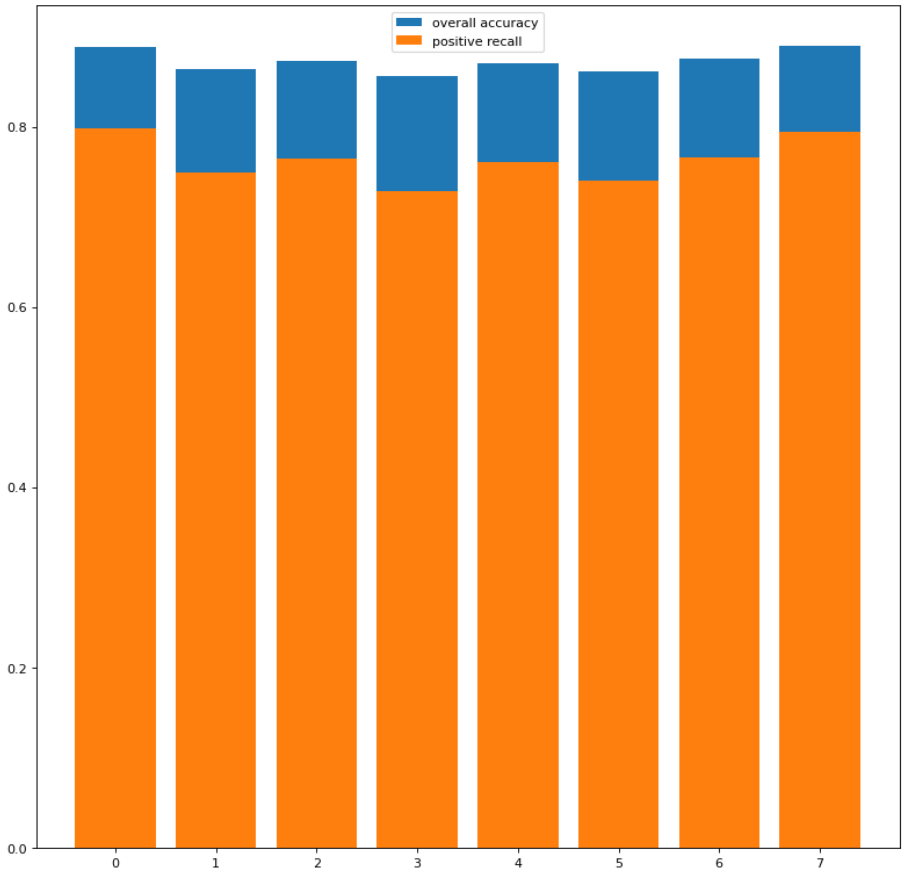
Accuracy	0.881155685
Precision	0.828312729
Positive recall	0.79177763
Negative recall	0.97053374



Here there is substantial improvement over the initial model, mostly in the positive recall percentage. There is a 18% increase in positive recall rate while only sacrificing 2% negative recall rate. This leads to an improvement in overall accuracy from 80% to 88%. This model would save a lot of time and man hours when it comes to trying to identify Information theory articles. Around $1/5^{th}$ of all positive cases are missed and $1/5^{th}$ of positively labelled cases are mislabelled.

4. Case study part 2

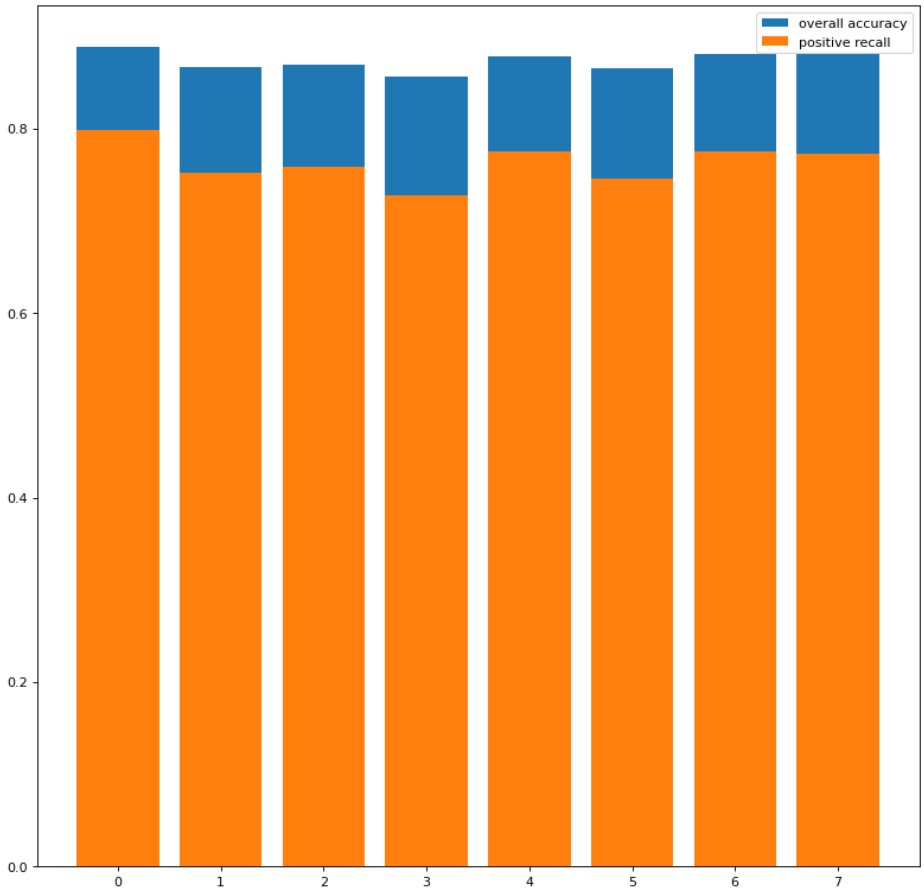
Question 1:
The dataset arxiv2019 was split into 8 separate timeframes and the articles with each timeframe used to evaluate model performance. Performance was measured using overall accuracy and positive case recall. This resulted in the following plot.



Question 2:
As time goes on the general trend for performance is decreasing. The only notable exception is the last period which breaks the trend slightly. The reason for the decrease in performance as time goes on is due to concept drift. The research fields are evolving and dealing with new kinds of research, as a result the vocab content of the articles will start to differ and this moves and changes the decision boundaries of the model.

Question 3 :
The method to be used would be a passive approach to combatting concept drift. The model will be updated every timestep using the previous timesteps information. Since in a business case we won't have the true labels for the data that we are to be classifying. This will be accomplished using the warm_start parameter from the RandomForestClassifier from sklearn.

The result of refitting the classifier every timestep is shown below:



The results from the passive concept drift mitigation approach have shown that the performance of the classifier does not degrade as sharply as it did before. However unrelated information may still be present, future versions can use a rolling window such that old data is no longer included in the training data. In this version all data available is used. The results here are more consistent than without the implementation.