Rabindra Khadka
DSSC, UNITS, Trieste

Detecting Important and Non Important Citation

1. Problem:

Not all citations are equally important, there is a need of deeper analysis of citation features to assess the impact of a research paper and identify a citation if it is important or non-important.

2: Background:

This project work mainly relies on the work presented by Valenzuela et al. (2015) [1] and classifies cited papers into important or non-important class. The features for this project were picked based on the best performing feature mentioned by David and Knoth [2] where the features used by Valenzuela et al are analysed and ranked. This is a supervised classification problem as both the features and the target for prediction are available. Random Forest Classifier was trained by providing features and target/labels. Given the unbalanced data set, precision and recall metrics were used for evaluating our model which stood at *40% and 60% respectively* along with ROC which stood at *AUROC of 0.84.*This matched pretty closely with the results presented by [1].

3. Approach:

Following workflow was adapted to tackle this problem:

1. Question the problem from different dynamics and determine the relevant data.
2. Access the data, pre-process and clean the data
3. Prepare the data by extracting features for the machine learning model.
4. Establish a baseline model to compare the performance of the model
5. Split data set into training and test set.
6. Perform hyper parameter tuning using cross validation
7. Evaluate and interpret the performance of model by looking at relevant metrics.

4. Data Acquisition:

Publicly accessible annotated data set provided by Valenzuela et al. was used for this project which is available at *https://allenai.org/data/data-all.html*. The data set consists of relevant and non-relevant human judgements of 465 tuples of citing and cited papers. Each citation was labelled '0' for Non-Important Class or '1' for Important Class. The data set has 85.4% citation tuples as incidental/non-important citations and 14.6% as important/influential citation.

For cleaning the data set, first all the pdfs of the citing and cited papers were collected as mentioned by Valenzuela in his work. These pdfs were converted to text using pdftotext in R. The pre-processing of the texts were carried out by using methods namely tokenization, stemming and lemmatization using NLTK, a python package. [4]

The final data set consists of following feature columns:

1. **Direct Citation in paper:** the number of citations in citing paper from cited paper.
2. **Abstract Similarity:** cosine similarity between the abstract of cited paper and citing paper.
3. **Class:** Target labelled as ‘0’ for non-important citation and ‘1’ as important citation.

Taking into account these above mentioned features, data were segregated into features and target variable. Then the data set was divided into training and test set at *a ratio of 7:3 with a random state of 42* after which the model was trained using scikit learn in python. In our project *the baseline is considered to be at 14.6%* important citations given by the annotated data provided by the work of Valenzuela et al.

5. Feature Extraction:

As identified in the work of Pride and Knoth [2], the feature namely direct citation and abstract similarity produced the best result. So, this project was performed by focusing on extracting these two features only and thus ignoring other features mentioned by [1].

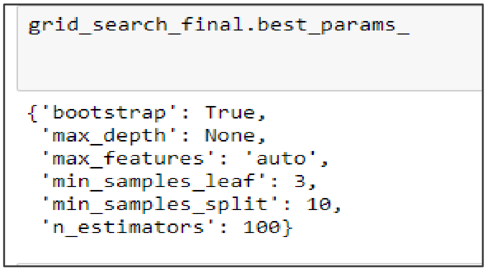
(a) Abstract Similarity: This abstract similarity was extracted by calculating *the cosine similarity* of *the tf-idf scores* obtained between abstracts of citing and cited paper. This process involves two steps, first converting each text into vector. *Term frequency (TF)* reflects the importance of the words based on their frequency in a document and *IDF* provides score to depict the frequency of a word across multiple documents. The higher the frequency of a word across documents, lower is its significance. [5] Second step involves computing the distance between two vectors which is based on the concept that cosine or dot product of same vectors is 1 and dissimilar vectors has dot product equal to 1.So the measure of similarity lies in between 0 and 1. [3]

Abstract similarity feature had the highest score of 0.63 between our two features which is supported by [2].

(b). Direct Citation: This feature received the score of 0.37 which resembles the score given by [1] at a rank of second. It was extracted by counting the total number of citations to the cited paper.

6. Tuning of Hyper parameters:

Hyper parameters were set before training our model. Some of the important parameters of Random Forest Classifier needed to be tuned are *the max depth of the forest, number of decision trees, max number of features to be considered for the split and min number of samples required at each node for split.* To avoid over fitting, cross validation method was employed



```
grid_search_final.best_params_

{'bootstrap': True,
 'max_depth': None,
 'max_features': 'auto',
 'min_samples_leaf': 3,
 'min_samples_split': 10,
 'n_estimators': 100}
```

Fig a: Best Parameters after tuning of parameters while tuning the hyper parameters. RandomizedsearchCV method from scikit learn helped to narrow down the search. For this project ***5 fold cross validation*** method was used for training the model. The GridSearchCV algorithm was used from scikit learn to find the optimal parameters for the classifier. The best parameter for the proposed model obtained can be seen in fig a.

7. Determining Performance Metrics:

After tuning of parameters involving CV for training the model, the test data set was used for making predictions. In order to evaluate the performance of our model which was trained on highly imbalance dataset; *Receiver Operating Characteristic curve* (ROC) along with *precision, recall and f1 score* analysis were observed.

7.1 ROC Analysis:

ROC curve gives us the overall performance of the classification model. The answer to the question how well our model detected important and non-important citation can be observed with ROC curve. The Random Forest classifier had the AUROC at 0.82 which resembled the number produced by Valenzuela et al. in their work. [1] This shows that with 82% of time the positive class i.e. important class is ranked higher than the non-important class. The model provided us a very upbeat result even with a small sample size and just two features.

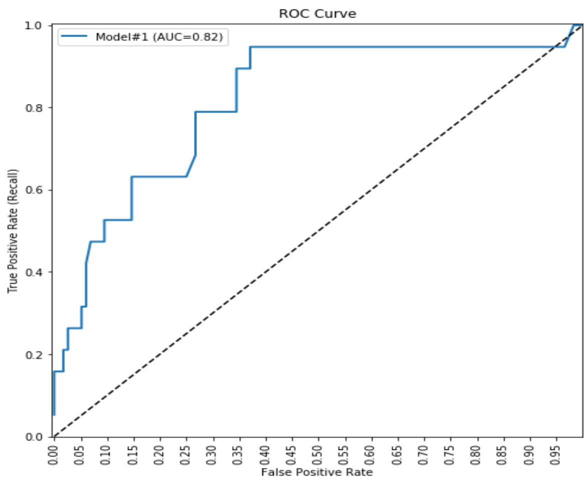


Fig b: ROC curve of Random Forest Classifier

7.2 Precision Recall Analysis:

The confusion matrix shown in table 1 shows us the four different prediction results which can be used to derive the Precision and Recall for a threshold of 0.30.

Actual	**Predicted**	
	Non Important Citation	Important Citation
Non Important Citation	**97 (TN)**	**18 (FP)**
Important Citation	**8(FN)**	**12(TP)**

Table 1: Confusion Matrix

The precision and recall can be calculated from the confusion matrix which provides a better picture in the case of imbalanced dataset like ours (85.4% incidental citations).In order to detect higher number of important citations, the model should be able to find the relevant data point. This can be done by trading off higher recall value with lower precision. We have the recall value of 60% given by [TP / (FN + TP)] and precision of 40%given by [TP / (FP + TP)] [6]. In other words, out of all truly relevant data i.e. important citations, 60 % of them are found by our classifier. When compared this with our baseline which is indicated by blue line in fig c, for the same recall score of 0.6, precision is below 0.2.Thus our model is clearly a better estimator with a good level of precision while having a modest recall score.

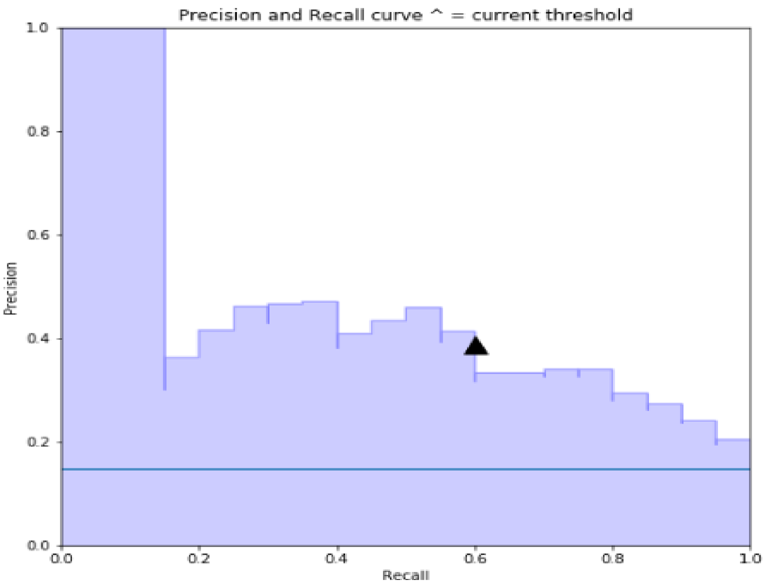


Fig c: Precision and Recall of Random Forest Classifier

8. Conclusion:

The proposed model yielded a good performance results that resembled with the previous work carried by [1]. So the results of the model shows that these two chosen features (abstract similarity and direct citation) are good indicators of the relevance or importance of a citation for a citing paper. The definition of citation category will greatly influence the result. The performance could be improved if higher sample size could be used and with extraction of other important features such as more frequently used words in the citing paper near the citation location.

Reference:

1. M. Valenzuela, V. Ha, and O. Etzioni. 2015. Identifying meaningful citations, *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.
2. Pride, David and Knoth, Petr (2017).Incidental or Influential? –A decade of using text-mining for citation function classification. In: 16th International Society of Scientometrics and Informetrics Conference, 16-20 Oct 2017, Wuhan.
3. https://www.nltk.org/book/ch06.html
4. https://learning.oreilly.com/library/view/python-text-processing/9781849513609/ch07s02.html
5. https://www.analyticsindiamag.com/nlp-case-study-identify/
6. https://www.oreilly.com/ideas/evaluating-machine-learning-models/page/3/evaluation-metrics