Scubed at 3C task A - A simple baseline for citation context purpose classification

Shubhanshu Mishra
shubhanshu.com
mishra@shubhanshu.com

Sudhanshu Mishra
Indian Institute of Technology Kanpur
Kanpur
India
sdhanshu@iitk.ac.in

Abstract

We present our team Scubed's approach in the 3C Citation Context Classification Task, Subtask A, citation context purpose classification. Our approach relies on text based features transformed via tf-idf features followed by training a variety of models which are capable of capturing non-linear features. Our best model on the leaderboard is a multi-layer perceptron which also performs best during our rerun. Our submission code for replicating experiments is at: https://github.com/napsternxg/Citation_Context_Classification.

1 Introduction

The number of research papers has increased exponentially in recent years. In order to efficiently access this scientific resource, we need automated solutions for extracting information from these records. Citations in research papers are important for multiple reasons e.g. comparing novelty (Mishra and Torvik, 2016), expertise (Mishra et al., 2018a), and self-citation patterns (Mishra et al., 2018b). For people new to the field, they are an important resource to increase their knowledge whereas for experts in the field they act as useful pointers to summarize the paper. Citations are also used to measure various indexes which showcase the influence and reach of the researchers in their field. However, these indexes give equal weight to each citation. It has been established that all citations are not equal (N. Kunnath et al., 2020; Mishra et al., 2018b). In many cases, cited papers are used as examples. Often, they are not influential to the paper itself.

In this paper we describe our team, Scubed's entry for the citation context purpose classification shared task (N. Kunnath et al., 2020). This work aims to develop models that can identify the purpose of citations in the research papers, and hence can then be used to produce better indexes and make research more easily accessible to everyone.

1.1 Related Work

There has been a significant amount of work done in this area to better understand the significance of citations in a paper (N. Kunnath et al., 2020). As the number of research papers increase with time, the algorithms for suggesting research papers become more and more important. These algorithms are a deciding factor for lots of measures of a researcher's influence in a field. The no. of citations of a paper are important for deciding measures such as h-index (Hirsch, 2005) and g-index (Egghe, 2006). These are influential measures for describing the significance of a researcher in a field. Scholars have argued that all of the citations in a paper should not have the same weight while determining the impact and reach of a paper. Moras et. al (Moravcsik and Murugesan, 1975) showed, that many references in research papers are redundant and quite often share little context with the citing paper. There have been many techniques for classifying citations as influential. However, one of the strongest baseline for this task is the prior citation count of the cited paper. Works of (Chubin and Moitra, 1975) show the effectiveness of citation count in determining influence. The work of (Zhu et al., 2015) points out suitable features for this task. They evaluated the performance of 5 classes of features, count, position, similarity, context and miscellaneous. They determined that counting the number of times a citation is referenced in a paper is the best estimator to determine the influence of a citation. (Hou et al., 2011) also showed that the count of a citation in a research paper is a simple and effective technique to assign its scientific contribution and influence. (Nazir et al., 2020) applied SVM, Random Forests and Kernel Linear Regression classifiers to identify important

and non-important citations. They used citation count and similarity scores using tf-idf features to train their models. Their results show that these techniques produce an improved precision score of 0.84 in these tasks.

2 Task and Data Description

This paper focuses on the WOSP 3C shared sub-task B. In this sub-task, we were required to classify the citation context in research papers on the basis of their influence and purpose in the paper. For this shared task we used the ACL-ARC dataset(Jurgens et al., 2018) . The dataset consisted of 3000 labeled data-points annotated using the ACT platform (Pride et al., 2019). The data provided contains the following fields:

- Unique Identifier
- COREID of Citing Paper
- Citing Paper Title
- Citing Paper Author
- Cited Paper Title
- Cited Paper Author
- Citation Context
- Citation Class Label
- Citation Influence Label

To identify the citation being considered a #AUTHORTAG is placed in the citation. For this task the Citation Class Label field was ignored. This was a multi-label classification task, where the following target labels were used :

- **BACKGROUND**
- **COMPARES_CONTRASTS**
- **EXTENSION**
- **FUTURE**
- **MOTIVATION**
- **USES**

To evaluate the models the macro-F1 score was used on the test data. The final score that was used to rank was not the public score but a different subset of data that was not visible to the participating teams. The teams were advised to make submissions that would perform the best overall and not just on the public subset.

3 Methodology

We utilize a simple approach based on text classification baseline methods. For the original submission we utilized a limited set of models. However, we trained additional models to conduct exhaustive evaluation for this paper. Below, we describe our workflow for pre-processing, feature extraction, and model-training.

3.1 Pre-Processing and Feature Extraction

The data provided was in raw text format which is not suitable for making predictions directly. In order to make useful predictions, it has to be first converted into numerical vector form that our models can process. The raw data consisted of columns having different attributes for which different feature extraction techniques had to be applied. For example, the *citing* and *cited title* consisted of a titles of the research papers whereas the *citation context* consisted of a description of the citation context. In order to efficiently process each column separately we used the *ColumnTransformer* module from the scikit-learn library (Pedregosa et al., 2011). Each of the column contained text data. To extract useful features from this text data we used the *TfidfVectorizer* from the scikit-learn (Pedregosa et al., 2011) library on each column. This generates the term frequency inverse document frequency(*tf-idf*) score for each of the texts in each column. The tf-idf score is a normalized count for the words occurring in the corpus. This type of feature however does not account for the position and inter-dependence of words. The tf-idf score is calculated as follows:

$$tf - idf(t, d) = tf(t, d) * idf(t) \tag{1}$$

$$idf(t) = \log\left(\frac{1+n}{1+df(t)}\right) + 1 \tag{2}$$

In the above equations, *tf* stands for term frequency which refers to the number of times a term *t* occurs in a document *d*. The *n* in **(2)** refers to the total number of documents present in the document set. *(Df(t))* refers to the document frequency which calculates the number of documents in the document set that contain the term *t*. The tf-idf score is a better feature compared to the count of words in a sentence. The tf-idf score down weights uninformative words like pronouns compared to more rare but informative words present in the document.

In the end we ended up using two version of text features for our models:

1. **Citing Context only (v1)**: uses only features extracted from *citation context* column. Our

hypothesis here is that citation context should have the highest signal for identifying how the citation is used.

2. **All features (v2)**: uses features extracted from *citation context* as well as *citing* and *cited title* column. Our hypothesis here is that using the combination of features from both citing and cited paper should improve the signal for identifying how the citation is used. However, we are also aware that this may also increase the proportion of noisy features.

3.2 Prediction Models

For this shared task we were allowed to submit a maximum of 5 models for evaluation on test data [1]. Our goal was to investigate usage of the most simple models based on proven linear and non-linear models which are faster and easier to train and deploy compared to the recent more powerful but resource hungry deep learning models. The following models were submitted for evaluation:

- **Logistic Regression Classifier (LR)**: A simple logistic regression model trained on the tf-idf features of 3 columns.
- **Random Forest (RF)**: Random Forest model with 100 trees in the forest and boot-strapping trained on the tf-idf features.
- **Gradient Boosting Classifier (GBT)**: A gradient boosted classifier with 100 boosting stages trained on the tf-idf features.
- **Multi-layer Perceptron Classifier (MLP)**: A 1 hidden layer multi-layer perceptron classifier with 100 nodes and Relu activation, optimized using Adam optimizer with a learning rate of 0.001 and momentum of 0.99.
- **Multi-layer Perceptron Classifier (MLP-3)**: A 3 hidden layer multi-layer perceptron classifier with 256, 256, and 128 nodes in the first, second and third layers with Relu activation optimized using Adam optimizer with a learning rate of 0.001 and momentum of 0.99.

All the models were trained using the scikit-learn library.

4 Results

Table 1 shows the the public and private leader board scores for each of our submissions for this task. Our MLP (v2) model performed best on the leader-board while similar to the top performing model (within 0.02 F1 score).

Table 1: Results for the Purpose Sub-task. 4* implies that according to the leader board our entry is better than the 4th position entry. The non-highlighted rankings are made on the basis of the leader board private scores visible to us.

S.No	Model	Private	Public	Rank
1	GBT	0.144	0.150	4*
2	RF2	0.144	0.142	4*
3	MLPC	0.182	0.176	3
6	**Best**	**0.206**	-	**1**

4.1 Replication model performance after leader board submission

After the final leader board ranking, we decided to replicate the model performance on the actual test set provided to us by the shared task organizers. Our evaluation scores may not match with the submitted solutions as the model changes on each run and we did not record the random seed for the original submission. This analysis was conducted to generate comparable results for all models across the training and test sets (see table 2), and to further inspect the performance of the model on each label (see table 3).

First, table 2 shows the evaluation scores of all the models on the test set. One consistent pattern emerges, v1 models which use only the citation context text as its feature, consistently perform much better than v2 models. Next, the best v1 as well as v2 models are MLP and MLP-3. It appears that inclusion of extra features leads to over-fitting which is also evident from the training evaluation scores.

Table 2: Model evaluation scores on the test data on retraining models after leader board ranking.

model	v1		v2	
	test	train	test	train
lr	0.135	0.296	0.120	0.281
rf	0.140	0.954	0.136	0.958
gbt	0.151	0.719	0.148	0.770
mlp-3	0.186	0.995	0.177	1.000
mlp	**0.187**	0.995	0.185	1.000

Second, in table 3 we investigate the per label

[1] https://www.kaggle.com/c/3c-shared-task-influence/rules

evaluation (in terms of F1 score) for each of the models. For both v1 and v2 features almost all models show similar performance on all labels. All models perform best on the Background label which is also the most frequent label. Overall, it appears that these baseline models are quite good at learning this task compared to other submissions, while being fast and easy to implement.

5 Discussion

Our results show that tradition tf-idf features give good performance for this shared task resulting in a strong baseline to compare against. Simple machine learning models like logistic regression, random forests, and gradient boosted trees perform well for this task but are superseeded by multi-layer prerceptron models. Furthermore, the citation context contains the maximum signal for predicting citation usage. We were able to achieve one of the top performances in the task within the number of submissions required in the task. Due to the small dataset, multiple submissions increase the likelihood of the models to over-fit to the test set. Furthermore, our methods show that deep learning methods (e.g. mlp and mlp-3) do give significant advantage over simpler machine learning methods. The minor loss in performance is acceptable compared to the increased speed and low computation of simple machine learning models.

Further analysis reveals that MLP based models are indeed over-fitting to the training data as shown by near perfect F1-score on the training data (see 2). Additionally, GBT models consistently achieve much better performance on the test set compared to other models, including RF model which was our best entry on the leader board. Furthermore, the highest performing label is the Influential label. All models (except LR) perform the worse on the Incidental when using all text features but when only using citation context, the label performance is similar across labels.

6 Conclusion

Our team 'Scubed' submitted 3 models for the citation context classification based on purpose task. Out of the submitted models the multi-layer perceptron classifier performed the best on the test set achieving third position in this task. This model gave a private score of **0.18146** on the test set. We were able to achieve competitive results under minimum trials using fast and computationally cheap machine learning models.

References

Daryl E. Chubin and Soumyo D. Moitra. 1975. Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science*, 5(4):423–441.

Leo Egghe. 2006. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152.

Jorge Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102:16569–72.

Wen-Ru Hou, Ming Li, and Deng-Ke Niu. 2011. Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution. *BioEssays*, 33(10):724–727.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association of Computational Linguistics*.

Shubhanshu Mishra, Brent D. Fegley, Jana Diesner, and Vetle I. Torvik. 2018a. Expertise as an aspect of author contributions. In *WORKSHOP ON INFORMETRIC AND SCIENTOMETRIC RESEARCH (SIG/MET)*, Vancouver.

Shubhanshu Mishra, Brent D. Fegley, Jana Diesner, and Vetle I. Torvik. 2018b. Self-citation is the hallmark of productive authors, of any gender. *PLOS ONE*, 13(9):e0195773.

Shubhanshu Mishra and Vetle I. Torvik. 2016. Quantifying Conceptual Novelty in the Biomedical Literature. *D-Lib magazine : the magazine of the Digital Library Forum*, 22(9-10).

Michael J. Moravcsik and Poovanalingam Murugesan. 1975. Some results on the function and quality of citations. *Social Studies of Science*, 5(1):86–92.

Suchetha N. Kunnath, David Pride, Bikash Gyawali, and Petr Knoth. 2020. Overview of the 2020 wosp 3c citation context classification task. In *Proceedings of The 8th International Workshop on Mining Scientific Publications, ACM/IEEE Joint Conference on Digital Libraries (JCDL 2020*, Wuhan, China.

Shahzad Nazir, Muhammad Asif, Shahbaz Ahmad, Faisal Bukhari, Muhammad Tanvir Afzal, and Hanan Aljuaid. 2020. Important citation identification by exploiting content and section-wise in-text citation count. *PLOS ONE*, 15(3):1–19.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos,

D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

D. Pride, P. Knoth, and J. Harag. 2019. Act: An annotation platform for citation typing at scale. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 329–330.

Xiaodan Zhu, Peter D. Turney, Daniel Lemire, and André Vellino. 2015. Measuring academic influence: Not all citations are equal. *CoRR*, abs/1501.06587.

Table 3: Per label model evaluation on the test data.

model	BACKGROUND	COMPARES CONTRASTS	EXTENSION	FUTURE	MOTIVATION	USES	accuracy	macro avg	weighted avg
Citing Context only (v1)									
lr	0.702	0.000	0.000	0.0	0.000	0.110	0.543	0.135	0.400
rf	0.692	0.042	0.032	0.0	0.018	0.058	0.528	0.140	0.396
gbt	0.683	0.057	0.056	0.0	0.000	0.110	0.518	0.151	0.400
mlp-3	0.641	0.202	0.022	0.0	0.031	0.219	0.467	0.186	0.412
mlp	0.639	0.206	0.049	0.0	0.028	0.198	0.462	0.187	0.410
All features (v2)									
lr	0.707	0.000	0.000	0.0	0.000	0.013	0.547	0.120	0.388
rf	0.698	0.075	0.000	0.0	0.000	0.045	0.535	0.136	0.397
gbt	0.700	0.071	0.000	0.0	0.000	0.114	0.534	0.148	0.408
mlp-3	0.663	0.175	0.000	0.0	0.059	0.165	0.492	0.177	0.414
mlp	0.649	0.176	0.065	0.0	0.060	0.163	0.478	0.185	0.411