Overview of the 2021 SDP 3C Citation Context Classification Shared Task

Suchetha N. Kunnath
KMi, The Open University
Milton Keynes, UK
snk56@open.ac.uk

David Pride
KMi, The Open University
Milton Keynes, UK
david.pride@open.ac.uk

Drahomira Herrmannova
Oak Ridge National Laboratory[*]
Oak Ridge, TN, US
herrmannovad@ornl.gov

Petr Knoth
KMi, The Open University
Milton Keynes, UK
petr.knoth@open.ac.uk

Abstract

This paper provides an overview of the 2021 3C Citation Context Classification shared task. The second edition of the shared task was organised as part of the 2nd Workshop on Scholarly Document Processing (SDP 2021). The task is composed of two subtasks: classifying citations based on their (Subtask A) *purpose* and (Subtask B) *influence*. As in the previous year, both tasks were hosted on Kaggle and used a portion of the new ACT dataset. A total of 22 teams participated in Subtask A, and 19 teams competed in Subtask B. All the participated systems were ranked based on their achieved macro f-score. The highest scores of 0.26973 and 0.60025 were reported for subtask A and B, respectively.

1 Introduction

Authors cite scholarly publications for a wide variety of reasons (Garfield, 1972). As a result, citations in a research paper cannot be considered the same. Hence, evaluating research requires the substitution of existing citation frequency based research assessment methods and adoption of new practises[1]. One such qualitative method that can be considered is to use the sentences surrounding the citation, known as citation context, for identifying the author's intent. Characterising citations by examining citation context from the citing paper to evaluate the cited paper has been proposed as a promising direction for research assessment by previous studies (Abu-Jbara et al., 2013; Teufel et al., 2006).

Towards this end, several methods and taxonomies have been devised by manually assessing and analysing samples of research papers. This includes the classification schemes introduced by Moravscik and Murugesan (Moravcsik and Murugesan, 1975), Chubin and Moitra (Chubin and Moitra, 1975), Spiegel-Rosing (Spiegel-Rosing, 1977). The low generalisability due to limited dataset sizes and practical difficulties in using the above mentioned earlier schemes resulted in developing new general purpose classification taxonomies. In a pioneering work by (Teufel et al., 2006), a new taxonomy consisting of 12 function classes has been created and used as part of an experiment to train models for automatic classification of citation function. Other prominent schemes developed in similar contexts include (Abu-Jbara et al., 2013; Jurgens et al., 2018; Cohan et al., 2019; Pride and Knoth, 2020).

Despite these efforts and more than a decade of research, it remains hard to compare and contrast existing solutions in this field due to the existence of multiple classification schemes and the absence of a benchmark suite. This is the primary motivation for introducing the 3C shared task. The goal of the 3C shared task is to analyse the text containing the citation, known as the citing sentence, in order to classify the reference anchor according to their purpose or function (Subtask A) and influence or impact (Subtask B). Following the last year, we used a fraction of the Academic Citation Typing (ACT) dataset (Pride et al., 2019; Pride and Knoth, 2020) for both the subtasks. We used the Kaggle In-

[*] This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

[1] https://sfdora.org/

Proceedings of the Second Workshop on Scholarly Document Processing, pages 150–158
June 10, 2021. ©2021 Association for Computational Linguistics

Task	Category	Label	Distribution	
			Trainset	Testset
Subtask A	**BACKGROUND**	**0**	54.9%	54.4%*
	COMPARES_CONTRASTS	**1**	12.27%	12.1%
	EXTENSION	**2**	5.70%	5.90%
	FUTURE	**3**	2.07%	1.50%
	MOTIVATION	**4**	9.20%	10.6%
	USES	**5**	15.83%	15.5%*
Subtask B	**INCIDENTAL**	**0**	52.3%	45.7%
	INFLUENTIAL	**1**	47.7%	54.3%

* Data distributions for the 1st 3C shared task were 54.6% and 15.3% respectively.

Table 1: Subtasks categories, labels and data distribution

Class competitions [2] environment to run the shared task, because we found it to be a highly suitable environment in the previous iteration of the 3C task in 2020. All the submitted systems were ranked using macro f-score.

This overview paper is organised as follows: Section 2 describes the related work; Section 3 discusses the shared task setup, the data used, the baselines, followed by task evaluation in Section 4. Section 5 summarises the participating system description. Section 6 and 7 present the results and the conclusion.

2 Related Work

Existing automated systems mostly use taxonomy and AI-based methods to identify the citation function and influence from the citation context. Some of the prevalent citation function classification schemes developed in the past use pre-defined categories, with number of classes ranging from small to medium to large (Teufel et al., 2006; Abu-Jbara et al., 2013; Jurgens et al., 2018; Cohan et al., 2019). Some of the function taxonomies also considered the polarity aspect of the citations (Abu-Jbara et al., 2013; Hernández-Alvarez and Gómez, 2015). All classification methods developed for influence detection uses a binary scheme (*Incidental/Non-Important/Non-influential* and *Influential/Important*) (Zhu et al., 2015; Valenzuela et al., 2015).

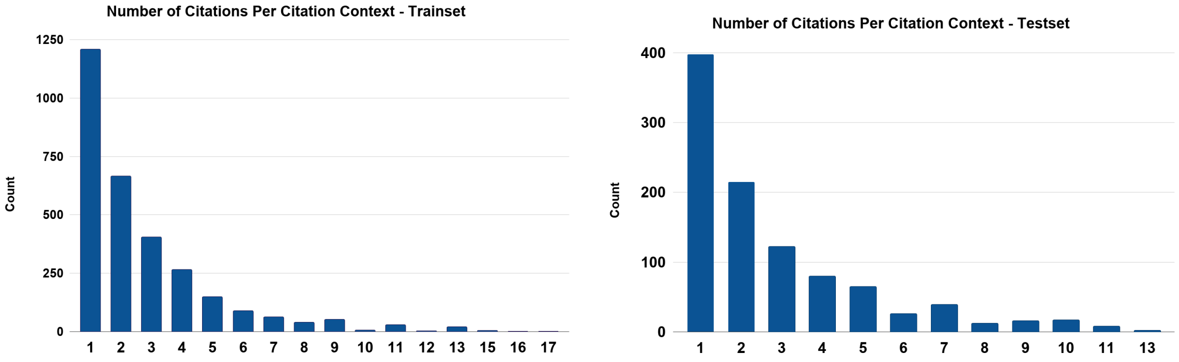
Earlier methods for citation classification used feature based supervised machine learning models like Random Forest (RF) (Jurgens et al., 2018; Pride and Knoth, 2017a,b; Valenzuela et al., 2015), Support Vector Machine (SVM) (Hernandez-Alvarez et al., 2017; Jha et al., 2017) etc., which requires the manual detection of the contextual and non-contextual features from the citation context, prior to training the model. With the development of larger datasets like SciCite (Cohan et al., 2019) and ACT (Pride and Knoth, 2020), more complex, transformer based models like SciBERT (Beltagy et al., 2019) were employed for solving this problem.

Date	Event
15/12/2020	Release of 1000 practise instances
26/02/2021	Kaggle competition start date + Release of Train, Test and Full text
30/04/2021	Kaggle competition end date

Table 2: 3C Shared task schedule

With the aim to enhance research in this field and to provide a general platform for the competing systems, we introduced the new 3C Citation Context Classification shared task in 2020 (Kunnath et al., 2020). Three teams participated in the citation purpose classification task and four teams contested in the influence task (Mishra and Mishra, 2020a,b; de Andrade and Gonçalves, 2020; Premjith and Soman, 2020). All the teams used simpler approaches such as TF-IDF for feature representation and other machine learning based models for classification. The highest score obtained were 0.205 (Subtask A) and 0.555 (Subtask B).

[2] https://www.kaggle.com/c/about/inclass



(a) Training set

(b) Test set

Figure 1: Distribution of the number of citations in the citation context

3 The 2021 3C Shared Task

The second version of the 3C shared task is organised by the researchers at the Knowledge Media Institute (KMi), The Open University, UK and Oak Ridge National Laboratory (ORNL), US. The 2021 3C shared task was part of the second Workshop on Scholarly Document Processing (SDP)[3], colocated with the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)[4].

3.1 Task Definition

The 3C shared task is a citation classification challenge having the following two subtasks:

- **Subtask A**: Multi-class classification of citations based on purpose with categories - *BACKGROUND*, *USES*, *COMPARES_CONTRASTS*, *MOTIVATION*, *EXTENSION*, and *FUTURE*.

- **Subtask B**: Binary classification of citations into *INCIDENTAL* or *INFLUENTIAL* classes, i.e. a task for identifying the importance of a citation.

Similar to previous year's shared task, we used the self-service Data Science competition hosting platform, Kaggle InClass competitions. Since Kaggle does not support hosting multiple tasks using a single interface, we used separate competitions for Subtask A `https://www.kaggle.com/c/3c-shared-task-purpose-v2/` and Subtask B `https://www.kaggle.com/c/3c-shared-task-influence-v2/`.

Table 2 shows the task schedule. The participating teams were required to comply with the following rules:

- Develop methods addressing the task and submit the results via Kaggle.

- Document and submit their method as a short system description paper.

- Provide source code for each method

3.2 Dataset

We used the same dataset from the previous year, with 3000 training instances. This year, we also provided the participants with the full text of all the papers present in both the train and testset. The full-text dump was extracted from the CORE[5](Knoth and Zdrahal, 2012) open access repository. We had to replace 12 out of 1000 instances from the previous year's testset due to the non-availability of a paper in the repository. Both the training and the testing datasets were in the csv format.

Table 1 shows the numeric values assigned for each of the citation purpose and influence labels. The table also contains the class distributions for both training and the testsets. The distribution of classes is similar to that used last year, except for a negligible variation in the testset. The highly imbalanced nature of the dataset makes the purpose classification a challenging task compared to Subtask B. Both datasets have the same 9 attributes as reported in (Kunnath et al., 2020), including the citation context, which represents the citing sentence. For each citing sentence, the citation considered

[3] `https://sdproc.org/2021/index.html`

[4] `https://2021.naacl.org/`

[5] `https://core.ac.uk/`

Team Name	**Leaderboard**	
	Public	**Private**
IREL	0.27968	**0.26973**
nlp_player	0.31385	0.26440
Duke Data Science	0.26160	0.26325
YDLP	0.28172	0.26229
YESNLP	0.31672	0.26189
IITP@3C	0.30258	0.26059
rookie	0.30042	0.25885
skating	0.32611	0.25775
CopeNLU	0.30027	0.25541
Xiaodong Wu	**0.33874**	0.25078
Test	0.29482	0.24650
jotline	0.31213	0.24231
Thunder Coder	0.27292	0.23784
Camp	0.26629	0.23476
VijayGao	0.23782	0.23147
admin	0.27325	0.22431
TakaComa	0.23046	0.22327
DLUFL	0.25730	0.22259
Anna Glazkova	0.24827	0.21948
Amrita_CEN_NLP	0.18369	0.21358
DT	0.22443	0.18891
Paul Larmuseau	0.18891	0.16884
majority_class_baseline	0.11938	0.11546

Table 3: Public and private leaderboard macro f1-scores for citation context classification based on purpose (Subtask A)

Team Name	**Leaderboard**	
	Public	**Private**
rookie	0.57832	**0.60025**
IREL	0.56523	0.59071
hello world	0.59427	0.59009
game	0.59835	0.57751
VijayGao	**0.60699**	0.57339
jotline	0.58861	0.57189
nlp_player	0.53427	0.56892
Xiaodong Wu	0.59012	0.56767
swltyy	0.57555	0.56702
cccxxxddd	0.52647	0.56621
CopeNLU	0.54633	0.56578
Duke Data Science	0.51596	0.55706
skating	0.59027	0.55582
YESNLP	0.58249	0.55500
Thunder Coder	0.56486	0.55015
Bhavyajeet Singh	0.55218	0.54712
IITP@3C	0.50396	0.53588
DLUFL	0.51302	0.49421
Amrita_CEN_NLP	0.54010	0.47516
majority_class_baseline	0.30362	0.32523

Table 4: Public and private leaderboard macro f1-scores for citation context classification based on influence (Subtask B)

is replaced with the tag, **"#AUTHOR_TAG"** as shown below:

"For example, previous studies have found that cross-institutional collaboration supports the diffusion of innovations and new ideas within a field (#AUTHOR_TAG and Darby, 1996)"

The dataset also included citation contexts with multiple citations, thus causing the tasks to be even more challenging. For instance, the below mentioned citation context has 6 citations in a single citing sentence:

*"Innovation through mimicry is likely to occur when innovations are socially visible (Mahajan and Peterson, 1985;Dos Santos and Peffers, 1998), when causes, conditions and consequences are known (absence of causal ambiguity, Barney, 1991;**#AUTHOR_TAG** and Venkatraman, 1992) and when the success of the innovation is unlikely to be determined by path dependencies (Barney, 1991;**#AUTHOR_TAG** and Venkatraman, 1992)"*

Figures 1a and 1b illustrates the distribution of number of citations in the citation context. 59.7% of instances in the training set has more than one citation in the citation context. Likewise, the majority of citation contexts in the testset has multiple citations (60.2%).

Another aspect is the multi-domain nature of the dataset used for this shared task. Unlike some of the existing datasets for purpose and influence classification tasks (Cohan et al., 2019; Jurgens et al., 2018; Valenzuela et al., 2015), which are mainly drawn from the Computer Science and Bio-Medical domains, the ACT dataset has papers from other domains as well. The following citation context from the dataset is derived from the Mathematics domain:

*"It can be considered as an infinite dimensional analogue of a completely integrable Hamiltonian system (**#AUTHOR_TAG** and Shabat 1972;Shabat 1976), where the Hamiltonian and the Poisson bracket is given by $t + (x), H = 0, (x), *(y) = i(xy), (1.2)H = 12\|x\|2\|\|\|4dx$(here * stands for the complex conjugate function)"*

The presence of symbols and equations in the citing sentence, as shown in the above example,

Team	Micro F1	Macro F1
IREL	0.503	**0.267**
Duke Data Science	**0.508**	0.259
IITP	0.474	0.256
Amrita_CEN_NLP	0.396	0.198

Table 5: Evaluation results for purpose classification task on the complete test set

causes extracting meaningful insights about the citations even more difficult.

3.3 The Baseline

As in the previous 3C shared task, the baseline submissions for both subtasks were based on the majority class: *BACKGROUND*, for Subtask A and *INCIDENTAL*, for Subtask B. The majority class baseline model received a public and private score of 0.11938 and 0.11546 for the purpose classification task. The scores obtained for the influence task are 0.30362 and 0.32523.

4 Evaluation

In order to rank the submitted results by the participating teams, we used macro f-score:

$$F1 - macro = \frac{1}{n}\sum_{i=1}^{n}\frac{2 \times _{Pi} \times _{Ri}}{_{Pi} + _{Ri}} \tag{1}$$

where $_{Pi}$ and $_{Ri}$ represents the precision and recall for class i and n is the number of classes.

The submission file has the following fields in the csv format: (1) Unique ID and (2) Citation Class/Influence label. To avoid possible over-fitting due to large number of submissions, we restricted the per day number of submissions to 5. The final evaluation is based on the private score obtained by the teams. The number of scores eligible for the final evaluation was limited to 5.

5 Participating System Description

This edition of the 3C Shared task witnessed the active participation of more teams compared to the last year's tasks. 22 teams participated in the purpose classification task, and 19 teams competed in the influence classification task. The following subsections describe some of the submitted systems in this shared task.

5.1 IREL

Team IREL[6] (Maheshwari et al., 2021) tested various machine learning and deep learning models and found out that BERT based models like SciBERT (Beltagy et al., 2019) and RoBERTa (Liu et al., 2019) outperformed Random Forest (RF). The best result was obtained for uncased SciBERT with a linear classification layer. The team finished as the winner for subtask A with a private macro f-score of 0.26973. Using a similar approach for subtask B, IREL achieved a score of 0.59071, thus becoming second. The team reported a drop in performance when using fields other than citation context as input to the models. To address the class imbalance problem, IREL used the weighted loss function.

5.2 Duke Data Science

The team Duke Data Science[7] (Oesterling et al., 2021) used a multi-tasking approach, inspired from the model by Cohen et al. (Cohan et al., 2019). Additionally, the team used hand-engineered features like citation frequency and position-based features, along with the TF-IDF, computed from the previous sentence, next sentence and the citation context, as input to the model. The team also made use of external sources like ACL-ARC (Jurgens et al., 2018) and datasets from (Cohan et al., 2019) for the auxiliary tasks to enhance the model performance. The best results obtained was 0.26325 and 0.26160 for the private and public testsets. However, the team also reported an improved private score of 0.28071, for post-evaluation submissions.

5.3 IITP@3C

Similar to Duke Data Science, the team IITP@3C[8] (Kaushik Varanasi et al., 2021) used auxiliary tasks to improve citation performance, following (Cohan et al., 2019), for the purpose classification task. They introduced a third scaffold task, the cited paper title prediction, besides the citation worthiness and section title, to understand the correlation between the citation context and the cited paper. Instead of the Glove-ELMo word embedding, IITP@3C used SciBERT for representing the citation context. For task B, however the team used a machine learning based approach using Random Forest (RF), along with the TF-IDF, citation subjectivity, similarity, self-citation and length based

[6] 10.6084/m9.figshare.14687298

[7] 10.6084/m9.figshare.14687307

[8] 10.6084/m9.figshare.14687319

Year	Subtask	#Teams	Private Score			
			Max	Min	Mean	SD
2021	Purpose	22	**26.97**	**16.88**	23.90	2.60
	Influence	19	**60.03**	**47.52**	55.84	3.05
2020	Purpose	3	**20.56**	**12.54**	17.08	4.11
	Influence	4	**55.57**	**51.53**	54.26	1.85

Table 6: Summary statistics for the private macro f-score values obtained for this year's and previous year's shared task. Max - maximum score, Min - minimum score, SD - Standard Deviation

features. The team reported the highest private scores of 0.26059 and 0.53588 for task A and B, respectively.

5.4 Amrita_CEN_NLP

Team Amrita_CEN_NLP[9] (Indhu S et al., 2021) experimented with Bi-directional Long Short Term Memory (BiLSTM) and RF for purpose and influence tasks. The team used not just the citation context but also cited title for both tasks. To mitigate the dataset skewness problem, a cost-sensitive learning approach was employed for RF and class weights for Bi-LSTM. The highest scores reported by the team were 0.0.21358 for Subtask A and 0.47516 for Subtask B, which are higher than the majority class baseline models.

6 Results

Table 3 outlines the public and private test scores obtained for all 22 teams, which participated in the citation purpose classification task. The highest and the lowest private scores on the leaderboard were 0.26973 and 0.16884. The top 6 teams scored greater than 0.26000. There is no significant difference between the scores, as indicated by the lower value of standard deviation (SD) in Table 6 for this year's purpose classification task. Table 5 shows the overall micro and macro f-scores on the entire test dataset for the purpose classification task. The table clearly signals higher values for micro f-scores when compared to the macro average scores, since the former focuses more on the majority class, unlike the latter, which treats all classes equally, despite the proportion.

For the influence classification task, the scores obtained for the 19 teams is given in Table 4. Team rookie obtained the highest private score of 0.60025, the only team to achieve a score greater than 0.6. Compared to the last year's score, as indicated in Table 6, there is considerable improvement in the performance (8.03% increase) of systems submitted by the teams for this subtask.

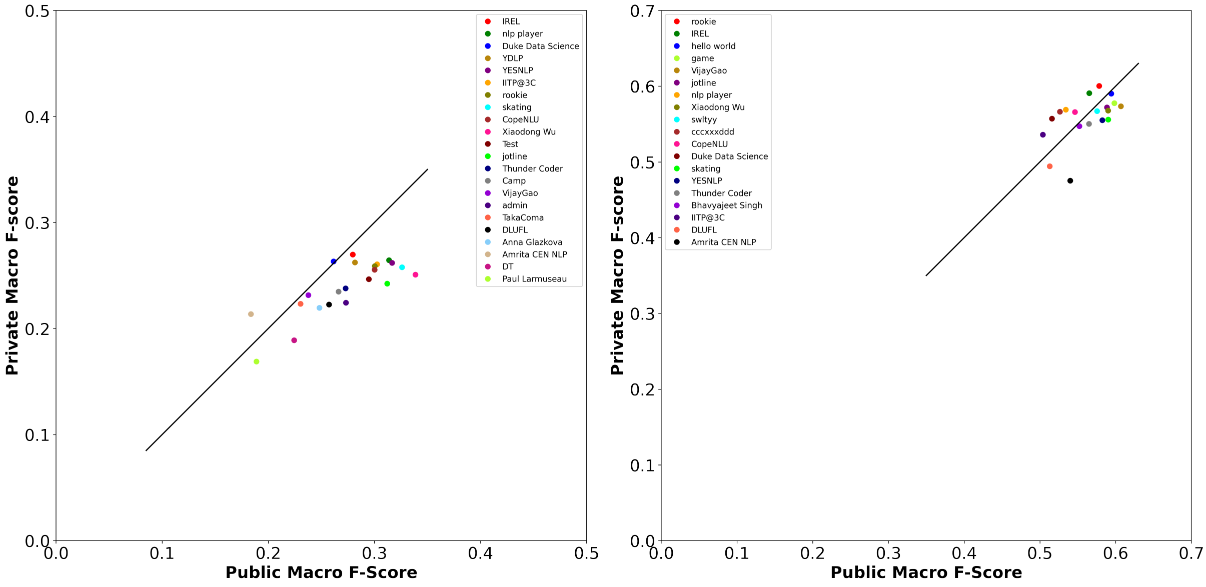
7 Discussion

The SDP 3C shared task is a continuation of last year's WOSP 3C shared task. The significant difference, however, is the inclusion of a full-text dataset for both subtasks. Another highlight of this year's shared task is the active participation of 27 teams from the start, out of which 14 of them competed in both subtasks in Kaggle. The overall scores obtained for both tasks shows considerable improvement when compared to the previous year's results (Table 6).

The paper and code submissions for the purpose classification tasks shows the use of deep learning based methods by all the teams, as opposed to the last year, when the successful teams relied primarily on more traditional non-deep machine learning methods. The winning team, IREL, used a pre-trained SciBERT model for predicting citation purpose. Interestingly, using just the citation context information, the team achieved the highest private score. Team Duke Data Science, which finished third on the leaderboard, however, used external datasets like ACL-ARC and SciCite as well for the purpose task. They also used the full-text dataset to extract the previous and next sentences for generating the TF-IDF. IITP@3C emphasised the need for exploiting additional information related to the cited paper. Like Duke Data Science, IITP@3C, too, used a similar state-of-the-art Bi-LSTM attention model and external dataset. The team Amrita_CEN_NLP, who competed in the previous competition, improved their results this year using a Bi-LSTM model.

For the influence classification task, we received

[9] 10.6084/m9.figshare.14687325



(a) Subtask A

(b) Subtask B

Figure 2: Public Vs Private Macro F-Score performance on the leaderboard

the paper and source code from three teams. Except for team IREL, both IITP@3C and Amrita_CEN_NLP used a feature-based machine learning approach. IITP@3C evaluated the feature importance and found that the similarity-based features computed between the citation context and cited/citing paper title resulted in increased performance. Amrita_CEN_NLP obtained the highest score with RF and fasttext embedding for the influence task. Team IREL, however, used the same SciBERT based model for this task too.

Figure 2 represents the public vs private macro f-scores plotted for both the subtasks. For the influence classification task, the graph 2b shows that the scores obtained by the teams are clustered more around the diagonal. However, for the purpose classification task, as shown in the graph 2b, there is significant deviation for the majority of points, indicating the possibility of over-fitting or under-fitting on the public data.

Although this year witnessed significantly more teams participating than in the previous shared task, despite our systematic encouragement, we only received research papers and source code submissions from four teams for Subtask A and and three teams for Subtask B. For the purpose classification task, out of the top five teams, only two teams submitted their methods. We did not receive the solution used by the top team on the leaderboard for the influence task. The open nature of the competition hosting platform resulted in the participation of several Kaggle enthusiasts, who competed with the other teams, without formally registering for the shared task.

8 Conclusion

This paper describes the overview of the 2nd 3C citation context classification shared task organised as part of the SDP workshop and run at NAACL. As in the previous year, two subtasks were organised to classify citations based on their purpose and influence. In addition to the act dataset, we provided the full-text data dump to the participants. This year, we have observed a substantial improvement in the private macro f-score values. We believe this is attributed to the use of advanced deep learning methods, the full-text dataset and external sources by the participating teams. However, the overall lower scores still indicate the need to address the challenges caused by the multi-domain nature of the dataset and the presence of more than one citation in the citation context for achieving better performing systems.

References

Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. 2013. Purpose and polarity of citation: Towards nlp-based bibliometrics. In *Proceedings of the 2013 conference of the North American chapter of the as-*

sociation for computational linguistics: Human language technologies, pages 596–606.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

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.

Arman Cohan, Waleed Ammar, Madeleine Van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. *arXiv preprint arXiv:1904.01608*.

Claudio Moisés Valiense de Andrade and Marcos André Gonçalves. 2020. Combining representations for effective citation classification. In *Proceedings of the 8th International Workshop on Mining Scientific Publications*, pages 54–58.

Eugene Garfield. 1972. Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479.

Myriam Hernández-Alvarez and José M Gómez. 2015. Citation impact categorization: for scientific literature. In *2015 IEEE 18th International Conference on Computational Science and Engineering*, pages 307–313. IEEE.

Myriam Hernandez-Alvarez, José M Gomez Soriano, and PATRICIO MARTÃNEZ-BARCO. 2017. Citation function, polarity and influence classification. *Natural Language Engineering*, 23(4):561.

Isha Indhu S, Kavya S Kumar, Lakshaya Karthikeyan, B Premjith, and K.P Soman. 2021. Amrita_cen_nlp@sdp2021 task a and b. In *Proceedings of the Second Workshop on Scholarly Document Processing*.

Rahul Jha, Amjad Abu-Jbara, Vahed Qazvinian, and Dragomir R Radev. 2017. Nlp-driven citation analysis for scientometrics. *Nat. Lang. Eng.*, 23(1):93–130.

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 for Computational Linguistics*, 6:391–406.

Kamal Kaushik Varanasi, Tirthankar Ghosal, Piyush Tiwary, and Muskaan Singh. 2021. Iitp-cuni@3c: Supervised approaches for citation classification (task a) and citation significance detection (task b). In *Proceedings of the Second Workshop on Scholarly Document Processing*.

Petr Knoth and Zdenek Zdrahal. 2012. Core: three access levels to underpin open access. *D-Lib Magazine*, 18(11/12):1–13.

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*, pages 75–83. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Himanshu Maheshwari, Bhavyajeet Singh, and Vasudeva Varma. 2021. Scibert sentence representation for citation context classification. In *Proceedings of the Second Workshop on Scholarly Document Processing*.

Shubhanshu Mishra and Sudhanshu Mishra. 2020a. Scubed at 3c task a - a simple baseline for citation context purpose classification. In *The 8th International Workshop on Mining Scientific Publications, ACM/IEEE Joint Conference on Digital Libraries (JCDL 2020)*, Virtual - China.

Shubhanshu Mishra and Sudhanshu Mishra. 2020b. Scubed at 3c task b - a simple baseline for citation context influence classification. In *The 8th International Workshop on Mining Scientific Publications, ACM/IEEE Joint Conference on Digital Libraries (JCDL 2020)*, Virtual - China.

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

Alex Oesterling, Angikar Ghosal, Haoyang Yu, Rui Xin, Yasa Baig, Lesia Semenova, and Cynthia Rudin. 2021. Multitask learning for citation purpose classification. In *Proceedings of the Second Workshop on Scholarly Document Processing*.

B Premjith and KP Soman. 2020. Amrita_cen_nlp@ wosp 3c citation context classification task. In *Proceedings of the 8th International Workshop on Mining Scientific Publications*, pages 71–74.

David Pride and Petr Knoth. 2017a. Incidental or influential?–a decade of using text-mining for citation function classification.

David Pride and Petr Knoth. 2017b. Incidental or influential?-challenges in automatically detecting citation importance using publication full texts. In *International conference on theory and practice of digital Libraries*, pages 572–578. Springer.

David Pride and Petr Knoth. 2020. An authoritative approach to citation classification. In *2020 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Virtual - China.

David Pride, Petr Knoth, and Jozef Harag. 2019. Act: an annotation platform for citation typing at scale. In *2019 ACM/IEEE Joint Conference on*

Digital Libraries (JCDL), pages 329–330, Urbana-Champaign, Illinois. IEEE.

Ina Spiegel-Rosing. 1977. Science studies: Bibliometric and content analysis. *Social Studies of Science*, 7(1):97–113.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 103–110.

Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *AAAI workshop: Scholarly big data*, volume 15, page 13.

Xiaodan Zhu, Peter Turney, Daniel Lemire, and André Vellino. 2015. Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, 66(2):408–427.