*International Journal of Information Science and Management*
*Vol. 24, No. 3, 2026, 89-104*
DOI: https://doi.org/10.22034/ijism.2026.2029231.1494 / DOR:

*Original Research*

# Advancing Multi-Task Learning Techniques for Citation Context Classification: Sentiment, Role Source, and Function

**Yaniasih Yaniasih**
PhD. in Computer Science, Research Center for Data and Information Science, National Research and Innovation Agency, Jakarta, Indonesia.
Corresponding Author: yaniasih@brin.go.id
ORCID iD: https://orcid.org/0000-0002-3389-6742

**Indra Budi**
Professor, Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia.
indra@cs.ui.ac.id
ORCID iD: https://orcid.org/0000-0002-2107-6552

Received: 21 May 2024
Reviewed: 02 June 2024
Accepted: 04 May 2026

## Abstract

The pattern of the in-text citation is significant in determining its value. This study focuses on advancing multi-task learning (MTL) techniques for citation-context classification: sentiment, role source, and function. Specifically, the research explores the integration of shared-trunk, cross-stitched, and shared-private networks with deep learning (DL) models, namely a convolutional neural network (CNN), a long short-term memory (LSTM), and a bidirectional LSTM. The dataset comprises 8566 sentences that include in-text citations of journal articles originating from Indonesia. To address the issue of unbalanced data, distinct loss weights were assigned to each output. Additionally, Bayesian optimization was employed to determine the hyperparameter sizes. The findings of the optimization process confirm that the learning rate is the most influential hyperparameter. The architectural complexity of the three MTL types varies, but this does not affect time efficiency, which is heavily influenced by the DL model employed. By achieving an F1 score of 0.87 for role source, 0.90 for function, and 0.81 for sentiment classification, the shared-private CNN model surpasses the current state-of-the-art model. The proposed model may improve citation analysis, making it beneficial for developing citation recommendation systems and conducting science evaluations.

**Keywords:** multi-task learning, deep learning, citation classification, hyperparameter optimization, imbalanced dataset, in-text citation

## Introduction

Classifying the context of citations in a text is essential to fully comprehending their meaning. The author's choice of language can reveal crucial information about whether the referenced article is genuinely significant or perfunctory. The level of significance can be determined by the author's positive, negative, or neutral evaluation of the cited article (Ghosh, Das & Chakraborty, 2018; Nazir et al., 2022). The author's purpose in referring to an article can also be seen in comparisons or explanations of results, as an introduction, or for use in technical

research (Yaniasih & Budi, 2021; Zhang, Zhao, Wang, Chen & Mahmood, 2022). In addition, citation analysis can identify the specific sections of an article being cited—such as the introduction, methods, or results—thereby providing greater insight into the contextual and functional dimensions of scholarly citations (Iqbal et al., 2021; Zhao, Feng, Luo, & Ye, 2019). Based on the three citation-context meanings discussed previously, the value of a citation in a scientific article can be evaluated more precisely.

The current approach to categorizing citation meanings is primarily conducted individually for each specific meaning using a single-task model. This approach has certain limitations, such as reliance on limited data and the substantial computational resources required (Yousif, Niu, Chambua & Khan, 2019). Multi-task learning (MTL) methods enable concurrent learning across multiple classification tasks, yielding several advantages. As stated by Crawshaw (2020), MTL has the potential to enhance data efficiency, particularly when the dataset is limited, mitigate overfitting through simultaneous training of representations, and accelerate training. Nevertheless, there is presently little literature that employs MTL to classify two meanings concurrently (Su, Prasad, Kan & Sugiyama, 2019; Yousif et al., 2019). Budi and Yaniasih (2023) are the only study we found that simultaneously classified three meanings using a multi-output model. However, they have not yet explored the various types of MTL models available. In contrast, this paper conducted a series of experiments to construct various MTL models, including shared-trunk, cross-stitched, and shared-private architectures. Additionally, this study distinguishes itself by addressing class imbalance through a class-weighting strategy, a technique not employed by Budi and Yaniasih (2023). Further advancements were made by conducting experiments on hyperparameter optimization. The experimentation with different MTL models, the strategies to mitigate data imbalance, and the hyperparameter optimization collectively constitute the key contributions of this study.

The three classification tasks in this study rely on shared and interconnected data. The MTL model is anticipated to yield superior classification outcomes compared to the single-task model. There exist several challenges in the MTL-based citation classification, which include the following:

- The analysis of citation patterns shows that minority classes have significantly less data but are highly important (Budi & Yaniasih, 2023), making their precise classification essential. However, addressing the imbalanced data through techniques such as resampling is impossible because the data consists of a single input with three output targets in the MTL model. On the other hand, the resampling technique only deals with a single input connected to a single output target.
- Challenges of hyper-parameter optimization in model construction arise due to the technical difficulty in applying existing optimization methods to multi-output models like MTL. This is primarily because conventional optimization is predicated upon a single loss value, and commonly utilized libraries lack support for MTL optimization.

Given those challenges, this study aims to construct MTL models that can effectively address three distinct meanings of citation sentences within a text. Furthermore, this study addresses techniques for resolving data imbalance and for optimizing hyperparameters in MTL models.

## Literature Review

MTL can be defined as $\{T_i\}_m^i$a model that $T_i$ performs a multitude of $m$ tasks that are interconnected yet distinguishable. The utilization of MTL has the potential to enhance learning by incorporating information from all $m$ tasks (Zhang & Yang, 2018). A literature review of multi-task models has identified three primary categories of MTL architecture, namely shared-trunk, cross-stitched, and shared-private. This study employed these three fundamental architectures.

The simplest model type in MTL is known as shared trunk (ST). ST is also referred to as a fully shared type (Liu, Qiu & Huang, 2017). This type represents a model wherein an initial layer is shared across all tasks, followed by distinct layers for each task's output. This particular model examines the data collectively and shares weights across tasks. The objective is to generalize information across various tasks. The expectation is that the models will effectively employ diverse knowledge representations to improve task classification accuracy. According to a survey by Crawshaw (2020), this model type has been widely used for various purposes in natural language processing and computer vision research. The model architecture has undergone significant evolution, with several variations in layer types and numbers, as well as in the hyperparameters used. Figure 1 illustrates the fundamental ST model architecture employed in this study.

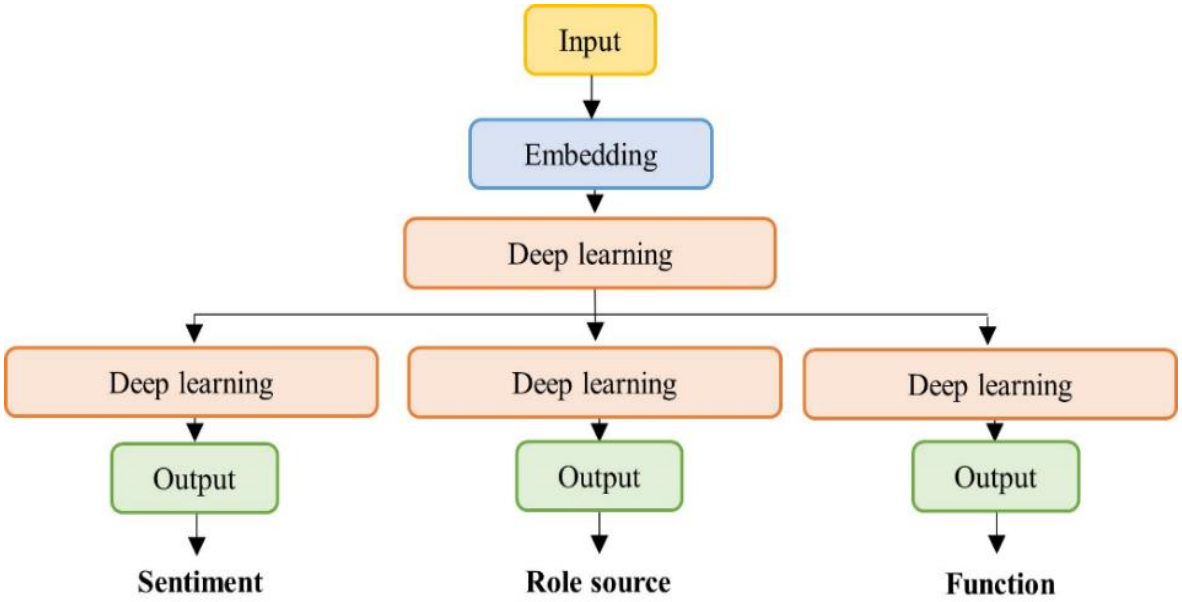



*Figure 1: Multi-task shared trunk model (MTL ST) (Modified from Tovedal, 2020)*

On the contrary, cross-stitched (CS) exhibits a distinguishing feature: a model comprising distinct networks for each task. Each layer's input in this model is derived from a linear combination of the previous layer's output from each task's network. The weights for each linear combination are learned and are unique to each task. Consequently, this enables each layer to selectively integrate task-specific information (Misra, Shrivastava, Gupta & Hebert, 2016; Ruder, 2017). It means that the model does not employ weight sharing across tasks. This approach allows the model to leverage information from multiple tasks while maintaining the confidentiality of the weights. An illustration of the CS MTL model architecture is shown in Figure 2.

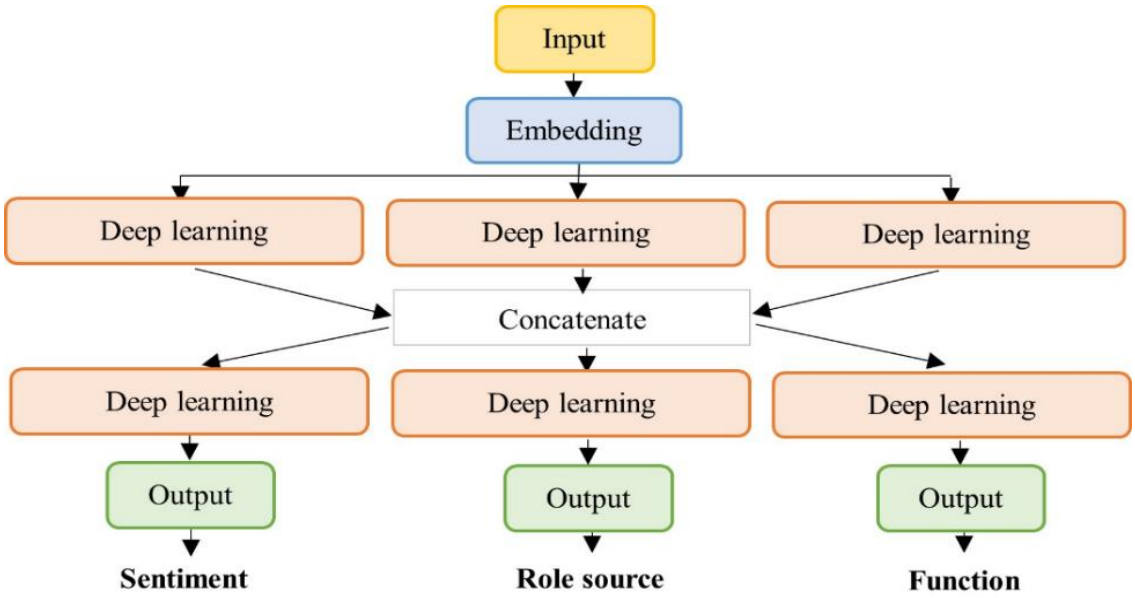



*Figure 2: Multi-task cross-stitched model (MTL CS*) (Modified from Tovedal, 2020)

Lastly, shared-private (SP) encompasses a model that incorporates both private and shared layers in parallel, with the private layer dedicated to maintaining task-specific features and the shared layer focused on preserving shared characteristics (Liu et al., 2017). SP has undergone rigorous experimentation in studies about the identification of named entities, the categorization of textual information, and the analysis of sequential patterns (Chen, Qiu, Liu & Huang, 2018; Tovedal, 2020; Wang, Lyu, Dong & Xu, 2019). The SP MTL model architecture employed is depicted in Figure 3.

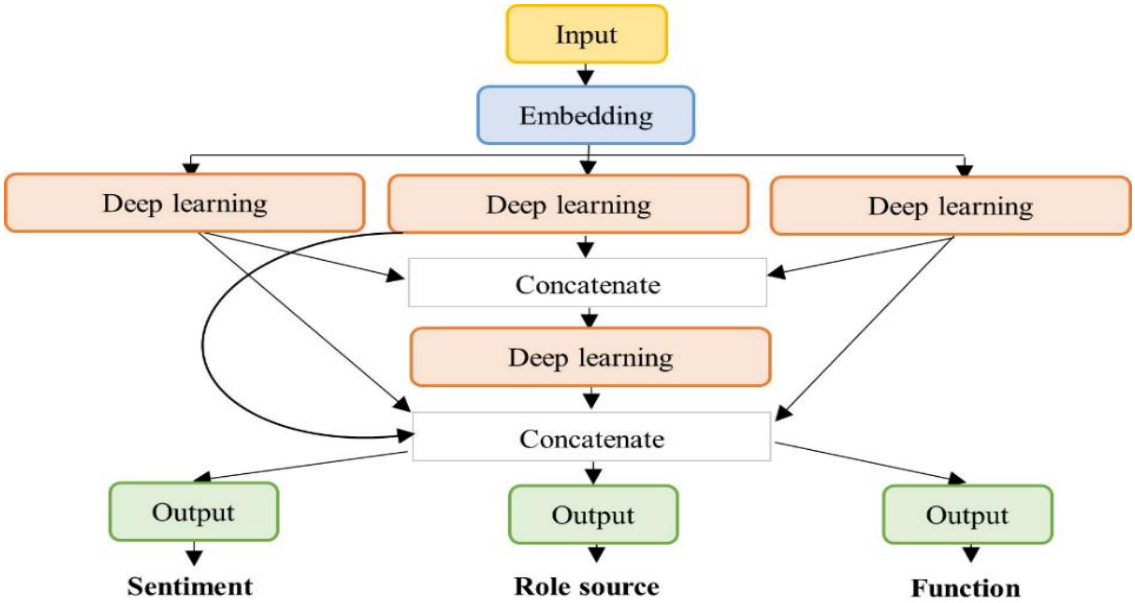



*Figure 3: Multi-task shared-private model (MTL SP) (modified from Tovedal, 2020)*

The MTL technique has been employed for two citation-meaning classifications, as depicted in Table 1. Several experiments with the MTL model for citation context classification exhibited superior, more efficient performance. The reviews conducted by Zhang and Yang

(2018) and Crawshaw (2020) also affirmed that numerous MTL studies have yielded favorable outcomes across a variety of task types, including those in natural language processing (NLP). When a model undertakes multiple tasks, it acquires a comprehensive representation of the data, rather than focusing solely on information that may be less significant and exclusive to a single task (referred to as data-dependent noise). This methodology helps mitigate overfitting in the model. The ability to effectively capture general information also gives the model an advantage in discerning relevant features from irrelevant ones. When confronted with a diverse array of data from multiple tasks, MTL can identify the most crucial features and apply a regularization function to the learned bias to enhance prediction accuracy (Ruder, 2017).

*Table 1*
*MTL-based citation classification*

| Literature | Objective | Data (number of sentences) | Method | Evaluation |
|---|---|---|---|---|
| Zhao, Luo, Feng, Zheng & Liu (2019) | Classified a new scheme of citation role and function | 3,008. Domain: computational linguistics, neural information, biomedical, and life sciences | The multi-task framework employed a BERT pre-trained model | F1 score 0.78, better than Logistic Regression, Support Vector Machine, CNN, Recurrent CNN, LSTM |
| Yousif et al. (2019) | Classified citation sentiment and purpose | Data 1: 3,568 and data 2: 1,768. Domain: Computer Science | MTL utilized CNN and Bidirectional (BiLSTM) | F1 score for sentiment 0.88, purpose 0.84 |
| Su et al. (2019) | Classified citation function and provenance | Sentiment data: 1,432; provenance data: 1,492. Domain: Computer Science | MTL utilized CNNs. | Accuracy for function 0.69 and provenance 0.79 |
| Cohan, Ammar, Van Zuylen, Van & Cady (2019) | Predicted citation function and location | Data 1: 1,941<br>Data 2: 11,020 | Structural scaffold, MTL used Glove and Elmo in the BiLSTM model | Data 1: F1 score: 0.67<br>Data 2: F1 score 0.84 |
| Baig et al. (2021) | Predicted citation function and location | 3,000 | TF-IDF embedding and LSTM with attention | >0.80 |

One of the key factors contributing to the success of MTL is the correlation between tasks. While there is no precise way to quantify this relationship, certain literature suggests that tasks are considered related if they utilize the same features for decision-making, share the same class hypothesis or learning bias, address the same problem, and are close in terms of parameter vectors (Ruder, 2017; Zhang & Yang, 2018). Sentiment classification and citation function are interconnected, enabling MTL utilization (Yousif et al., 2019). Similarly, there are connections among role and function (Zhao, Luo, Feng, Zheng & Liu, 2019), function and location (Cohan et al., 2019; Baig et al., 2021), and function and provenance (Su et al., 2019). Based on this

information, it can be assumed that all the above citation variables are related, and experiments can be conducted using the MTL approach for all variables.

Recent advancements in the literature indicate that the MTL approach outperforms single-task models in classifying citation meaning. However, the application of MTL has predominantly been limited to two tasks, with the most commonly used MTL type being the fully shared-layer model. Other MTL techniques, such as cross-stitched and shared-private models, have yet to be thoroughly explored. Notably, different MTL approaches can yield varying performances on the same dataset. This article addresses these gaps by classifying three citation-meaning tasks using three different types of MTL models. The findings are anticipated to contribute to the literature on citation meaning analysis in texts by applying artificial intelligence and natural language processing methodologies.

## Materials and Methods

### Dataset

To date, only one study has examined the three meanings of citation: sentiment, role source, and function (Budi & Yaniasih, 2023). Hence, this investigation employed the identical dataset utilized in their research. Using the same dataset facilitates comparison of experimental outcomes and obviates the need to build a dataset from scratch. The data utilized in this study are presented in Table 2. The aforementioned dataset comprises 8,566 citation-context sentences from 29 journals in Indonesia.

*Table 2*
*The statistical dataset*

| Citation meaning | Label class | Number of citations | Percentage |
|---|---|---|---|
| Sentiment | Positive sentiment | 422 | 2.47% |
| | Neutral sentiment | 7,932 | 4.93% |
| | Negative sentiment | 212 | 92.60% |
| Role source | Data role source | 32 | 0.37% |
| | Method role source | 1,347 | 15.72% |
| | Result role source | 1,316 | 15.36% |
| | Supplement role source | 5,871 | 68.54% |
| Function | Introduction function | 3,730 | 5.10% |
| | Explanation function | 645 | 7.53% |
| | Utilization function | 692 | 8.08% |
| | Related function | 3,062 | 35.75% |
| | Comparison function | 437 | 43.54% |

*(source: Budi and Yaniasih, 2023)*

### Model design

The model development stage involved preparing a basic multi-task learning model, optimizing hyperparameters, and evaluating the optimized model. The basic model utilized CNN, LSTM, and BiLSTM architectures. The CNN model was comprised of input, embedding, CNN, max-pooling, flattening, dense, dropout, and output layers. The LSTM and Bi-LSTM models consisted of input, embedding, LSTM, flattening, dense, dropout, and output layers.

Each layer of the model has a hyperparameter that serves as a measure. The size of these hyperparameters greatly influences the performance of a deep learning (DL) model. However, determining the size must be done at the beginning, as it cannot adapt to data characteristics such as parameter count. Hence, an optimization process for hyperparameters is necessary. This process typically involves experimenting with various combinations of hyperparameter sizes within the model. Traditional methods like grid search and random search are inefficient for deep learning MTL models due to the large number of combinations, which require significant time for experimental trials. This study explored several tools for hyperparameter optimization and found that a Bayesian optimization tool, namely Optuna (Akiba, Sano, Yanase, Ohta & Koyama, 2019), best aligned with the research objectives. The available optimization size options were as follows:

- Embedding size = 32, 64, 128
- Filter size = 32, 64, 128
- Kernel size = 3, 4, 5
- Dense unit = 32, 64, 128
- Dropout rate = 0, 0.25, 0.5, 0.75
- Learning rate = 1, 0.1, 0.01, 0.001, 0.0001, 0.0001
- Batch size = 16, 32, 64

The results chapter presents an analysis of the hyperparameter optimization results.

The optimized model was then used to classify the data set. The loss in the training data was calculated using a 10-fold cross-validation scheme with class weights to address limited and unbalanced data. To evaluate the model's performance, the following metrics were used: accuracy, precision, recall, and macro F1 score (Lever, Krzywinski & Altman, 2016). The performance was then compared to that of the best model produced by Budi and Yaniasih (2023). Additionally, the execution times of the models were also compared. Execution time is important in model development because it relates to resource requirements and computational costs (Justus, Brennan, Bonner & McGough, 2018).

## Results

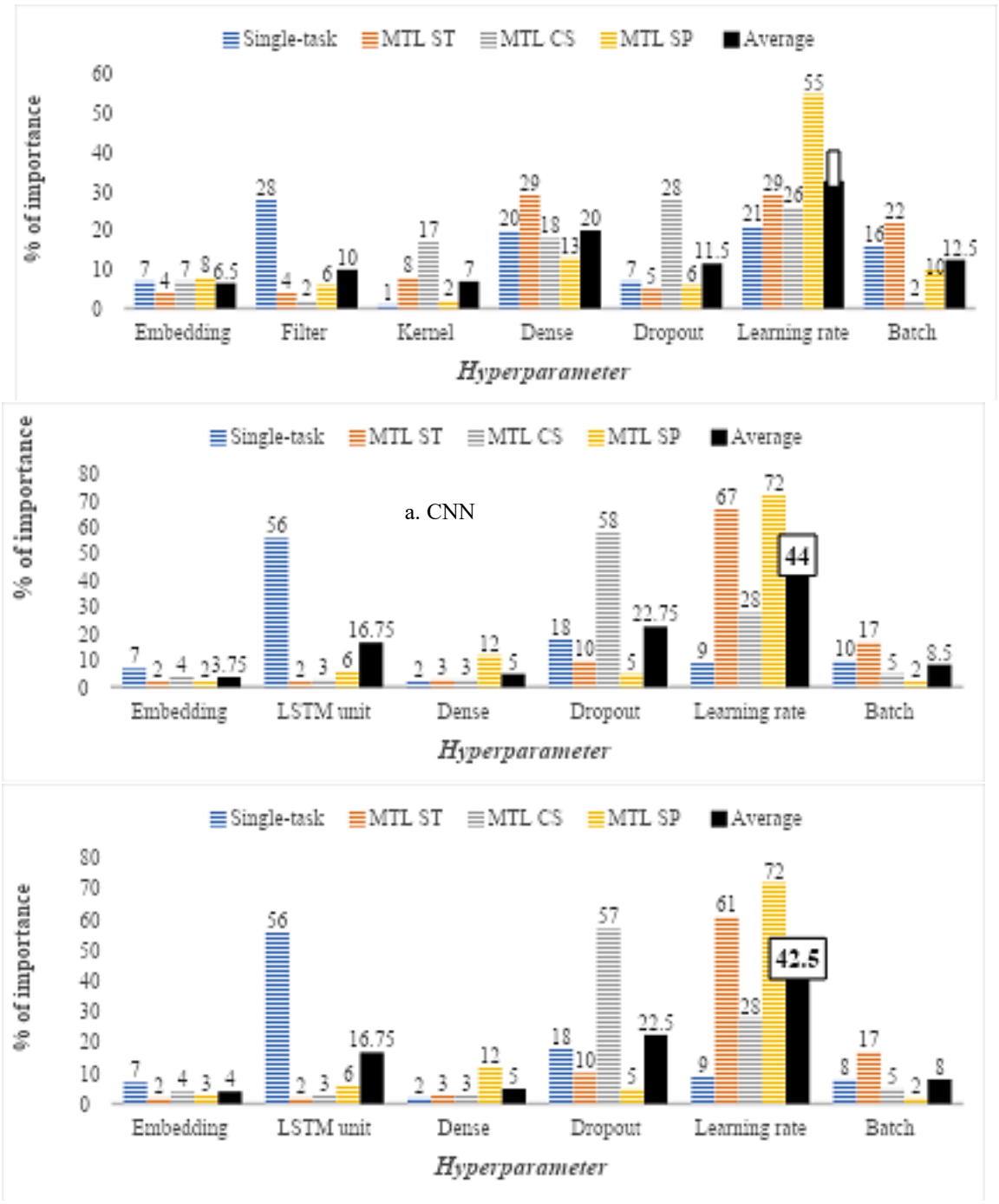
### Hyperparameter optimization

Each model has a hyperparameter size, and the optimization process seeks values that yield the best performance. The optimal size obtained from hyperparameter optimization for each model is shown in Table 3. The optimal values of each hyperparameter varied across models and architectures. To illustrate, for word embedding size, three CNN-based models attained optimal performance at 128, whereas the MTL ST model attained optimal performance at 32. Similarly, for the same hyperparameter, the LSTM architecture and BiLSTM yielded the same value: 32 for the MTL ST model, 64 for the single-task and MTL SP models, and 128 for the MTL CS model. Among all hyperparameters, the learning rate had the most consistent value. Nearly all architectures and models achieved optimal values at a learning rate of 0.001, except for two models: LSTM MTL SP and BiLSTM MTP SP.

*Table 3*
*The optimal hyperparameter optimization value*

| Architecture | Hyperparameter | The optimal size for each model | | |
|---|---|---|---|---|
| | | MTL ST | MTL CS | MTL SP |
| CNN | Embedding | 32 | 128 | 128 |
| | Filter | 32 | 128 | 64 |
| | Kernel | 4 | 3 | 5 |
| | Dense | 64 | 32 | 128 |
| | Dropout | 0 | 0,5 | 0,5 |
| | Learning rate | 0.001 | 0.001 | 0.001 |
| | Batch | 16 | 16 | 16 |
| LSTM | Embedding | 32 | 128 | 64 |
| | LSTM unit | 32 | 32 | 128 |
| | Dense | 128 | 64 | 64 |
| | Dropout | 0,25 | 0 | 0 |
| | Learning rate | 0.001 | 0.001 | 0.01 |
| | Batch | 32 | 64 | 64 |
| BiLSTM | Embedding | 32 | 128 | 64 |
| | LSTM unit | 32 | 32 | 128 |
| | Dense | 128 | 64 | 64 |
| | Dropout | 25 | 0 | 0 |
| | Learning rate | 0.001 | 0.001 | 0.01 |
| | Batch | 32 | 64 | 64 |

Analysis of the optimization results reveals the importance of each hyperparameter. Figure 4 illustrates the percentage importance of each hyperparameter. In the CNN architecture, the learning rate for MTL SP was the most important, at 55%. In single-task and MTL CNN models, the learning rate accounted for over 20%. The average importance of the learning rate was 32.5%, the highest among the hyperparameters. The dense hyperparameter also exhibited considerable importance across all types, exceeding 13%. On the other hand, the importance of other hyperparameters, such as filter and dropout, varied in value. In the single-task model, their importance was 28%, whereas in MTL CS, it was very low at 2%. In LSTM and BiLSTM, the importance of hyperparameters also makes the learning rate the most significant. The average importance level was even higher than that of CNN, reaching 44% for LSTM and 42.5% for BiLSTM. In terms of learning rate, the order of importance in these two architectures was dropout, LSTM units, batch, dense, and embedding, with the latter being the least important.




Bi LSTM

*Figure 4: Importance level of hyperparameters on model performance*

## Model performance

The sentiment classification experiment results are displayed in Table 4. All models achieved an accuracy of ≥ 0.90 in sentiment classification. The multi-task share-private model (MTL SP) CNN achieved the highest accuracy of 0.97, while the single-output CNN model using n-gram vector features had the lowest accuracy. The high accuracy values in all models may be attributed to the significant class imbalance in the sentiment dataset, with the neutral class comprising 92.60% of the data. Accuracy is the percentage of correct predictions made

by the model out of all predictions. The accuracy figures are heavily influenced by the proportion of the majority class.

To assess the quality of citations, it is crucial to have a model that can effectively detect minority sentiment classes, such as positive and negative sentiment. This is why the analysis of citation sentiment requires a model that excels in detecting these classes. To determine whether a model can detect these classes effectively, precision and recall are key metrics. The precision values of all models were quite diverse. The MTL ST model, which used LSTM and BiLSTM, achieved the lowest precision of 0.52, while the MTL SP CNN model achieved the highest precision of 0.89. On the other hand, most models had very low recall, less than 0.60. However, three models, namely single-output CNN, MTL CS CNN, and MTL SP CNN, achieved a recall higher than 0.60. Among these three models, the single-output CNN model had the highest recall, 0.87. The low recall across most models resulted in low F1, with only two models achieving high recall and high F1. The single-output CNN model achieved the highest F1 score of 0.85, followed by the MTL SP CNN model at 0.81.

*Table 4*
*Performance comparison of sentiment classification models*

| Model | Accuracy | Precision | Recall | F1 score | Time |
|---|---|---|---|---|---|
| Single-output CNN | 0.91 | 0.86 | **0.87** | **0.85** | 25 m 8 d |
| Multi-task ST LSTM | 0.94 | 0.52 | 0.46 | 0.48 | 30m 39d |
| Multi-task ST BiLSTM | 0.94 | 0.52 | 0.50 | 0.51 | 30m53d |
| Multi-task ST CNN | 0.94 | 0.65 | 0.47 | 0.49 | 3m29d |
| Multi-task CS LSTM | 0.94 | 0.64 | 0.47 | 0.51 | 27m24d |
| Multi-task CS BiLSTM | 0.95 | 0.68 | 0.55 | 0.60 | 3j33m48 |
| Multi-task CS CNN | 0.96 | 0.85 | 0.69 | 0.75 | 1j7m39d |
| Multi-task SP LSTM | 0.94 | 0.64 | 0.49 | 0.53 | 40 m 24 d |
| Multi-task SP BiLSTM | 0.95 | 0.85 | 0.55 | 0.57 | 1j 10m 19d |
| Multi-task SP CNN | **0.97** | **0.89** | 0.75 | 0.81 | 37m11d |

*Bold indicates the highest value*

Citation role source classification experiments are depicted in Table 5. The role classification used a more balanced dataset and achieved superior performance compared to sentiment classification. The accuracy values, ranging from 0.77 to 0.97, were lower than the average accuracy value for sentiment classification. However, all models exhibited precision values greater than or equal to 0.70 and recall values greater than 0.60, which were significantly better and more uniform than those for sentiment classification. The F1 score for all models was greater than or equal to 0.70, with the MTL SP CNN model attaining the highest value of 0.87. Furthermore, this model achieved the highest recall, 0.84, and a commendable precision, 0.93, though not the highest. Other models that gained an F1 score exceeding 0.80 included MTL SP CNN and MTL CS CNN (0.86), as well as MTL ST CNN (0.84).

*Table 5*

*Performance comparison of role source classification models*

| Model | Accuracy | Precision | Recall | F1 score | Time |
|---|---|---|---|---|---|
| Multi-output CNN | 0.96 | 0.94 | 0.77 | 0.84 | 25 m 8 d |
| Multi-task ST LSTM | 0.91 | 0.66 | 0.64 | 0.65 | 30m 39d |
| Multi-task ST BiLSTM | 0.90 | 0.66 | 0.64 | 0.65 | 30m53d |
| Multi-task ST CNN | 0.96 | 0.94 | 0.77 | 0.84 | 3m29d |
| Multi-task CS LSTM | 0.95 | 0.70 | 0.70 | 0.70 | 27m24d |
| Multi-task CS BiLSTM | 0.96 | **0.97** | 0.74 | 0.77 | 3j33m33d |
| Multi-task CS CNN | **0.97** | **0.97** | 0.82 | 0.86 | 1j7m39d |
| Multi-task SP LSTM | 0.95 | 0.94 | 0.71 | 0.72 | 40 m 24 d |
| Multi-task SP BiLSTM | 0.96 | 0.93 | 0.82 | 0.86 | 30m53d |
| Multi-task SP CNN | 0.94 | 0.93 | **0.84** | **0.87** | 37m11d |

*Bold indicates the highest value*

Table 6 reveals the results of the experiments conducted on the classification of citation functions. It was evident from these results that the performance of MTL models surpassed that of sentiment and role source classification. The model exhibited a range of accuracy, with the lowest being 0.86 and the highest being 0.93. Similarly, the precision ranges from 0.83 to 0.92, the recall from 0.78 to 0.90, and the F1 score from 0.80 to 0.90. Notably, the MTL SP CNN model achieved the highest F1 score and the highest recall.

*Table 6*
*Performance comparison of function classification models*

| Model | Accuracy | Precision | Recall | F1 score | Time |
|---|---|---|---|---|---|
| Multi-output CNN | 0.91 | 0.87 | 0.89 | 0.88 | 25 m 8 d |
| Multi-task ST LSTM | 0.86 | 0.87 | 0.80 | 0.82 | 30m 39d |
| Multi-task ST BiLSTM | **0.92** | **0.92** | 0.87 | 0.87 | 30m53d |
| Multi-task ST CNN | 0.91 | 0.87 | 0.89 | 0.88 | 3m29d |
| Multi-task CS LSTM | 0.90 | 0.90 | 0.88 | 0.89 | 27m24d |
| Multi-task CS BiLSTM | 0.91 | 0.90 | 0.86 | 0.88 | 3j33m48d |
| Multi-task CS CNN | 0.88 | 0.83 | 0.78 | 0.80 | 1j7m39d |
| Multi-task SP LSTM | **0.92** | 0.87 | 0.78 | 0.80 | 40 m 24 d |
| Multi-task SP BiLSTM | 0.90 | **0.92** | 0.87 | 0.87 | 30m53d |
| Multi-task SP CNN | **0.92** | 0.91 | **0.90** | **0.90** | 37m11d |

*Bold indicates the highest value*

## Discussion

Discussions of the hyperparameters that most strongly influence model performance, particularly in MTL, have not been extensively investigated. Two studies have been conducted to examine the significance of hyperparameters in DL. The significance of hyperparameters depends heavily on the model and data. However, research employing various datasets reveals that the two hyperparameters with the greatest impact on DL model performance are the learning rate and the number of epochs (Sharma, Van Rijn, Hutter & Müller, 2019). The findings of this study, which prioritize the learning rate as the most crucial hyperparameter, align with this research. Nevertheless, the results from the LSTM model differ from those of Sewak, Sahay & Rathore (2020), who identified sequence length, number of layers, and

activation function as the most sensitive hyperparameters for LSTM performance. This discrepancy may be attributed to Sewak's research, which does not examine the learning rate. Additionally, disparities between the datasets and models used further contribute to differences in findings.

Determining the optimal model to simultaneously classify three citation meanings requires careful consideration of its efficiency and effectiveness. One of the primary objectives of the multi-task model is to achieve time efficiency. An analysis of training time efficiency across models reveals no discernible pattern among models and architectures. The ST and SP models exhibited similar training times, whereas the MTL CS type required significantly longer training than the other models. Generally, in terms of time efficiency, the CNN model utilizing MTL ST emerged as the most efficient. Following closely behind were several models, namely single-output CNN, MTL ST, SP BiLSTM, and MTL SP CNN, which exhibit nearly identical time requirements.

The results from the sentiment classification demonstrate that both single-task and multi-task models generally achieve commendable accuracy. However, all the precision values were comparatively lower, and the recall values were significantly lower. This observation implies that although the constructed models achieved high accuracy in identifying true positives, they tended to misidentify a substantial number of false negatives, indicating lower sensitivity. The model from previous research, the single-output CNN, exhibited exceptional accuracy, precision, and recall, surpassing a threshold of 0.85. Consequently, it obtained the highest F1 score. One of the MTL models, MTL SP CNN, also displayed satisfactory performance, with improved accuracy and precision compared to the single-output CNN. However, its recall value was inferior.

The analysis of classification outcomes for three citation meanings demonstrates that both single- and multi-task CNN models outperform MTL models with LSTM and BiLSTM architectures. The sentiment classification task yielded the best performance with the single-output CNN model developed by Budi and Yaniasih (2023). The MTL SP CNN model achieved the highest F1 score for role and function classification, although it ranked second for sentiment classification. Similar to the study by Budi and Yaniasih (2023), the multi-output model, including the multi-task model, demonstrates better performance in classifying roles and functions but falls short of surpassing the single-output model in sentiment classification. Nevertheless, given the research objective of developing a model capable of simultaneously classifying three citation meanings, the MTL SP CNN emerges as the best-performing model. In this research, SP models outperform ST and CS models. This finding is attributable to the SP model's assimilation of knowledge, which comprises both task-specific information and overarching insights derived from all tasks. Consequently, such models exhibit superior generalization capabilities.

The best model, MTL SP CNN, has accomplished the objectives of this study. However, the methodology employed in this study still has several weaknesses. Firstly, the scope and quantity of data are limited to Indonesian journals. This limitation is problematic as citation patterns are influenced by culture, scientific field, and other factors. Nonetheless, it also offers benefits, as citation research has traditionally been dominated by data from international journals in affluent countries. The findings of this research contribute to enhancing the citation landscape of non-developed countries. Secondly, the classification outcomes for the sentiment

class still favor the single-output model. Further investigation is required to develop a multi-task model that effectively classifies sentiment classes.

The sentiment, role source, and function meanings of citation sentences reflect the significance of citations. However, the extent to which these factors should influence the assessment of citation value remains unexplored. The measurement of citation value based on sentiment meaning has been addressed by Ghosh et al. (2018), who proposed the M-index, and Nicholson et al. (2021), who introduced the Scite Journal Index (SJI). Both indices assess citation value according to sentiment categories, with positive sentiment assigned a higher value than neutral and negative sentiment a lower value than neutral.

The evaluation of role, source, and function meanings in citation assessments can draw on the citation investigation parameters proposed by Swales (1986). For example, a role source categorized as data and method typically indicates an operational relationship between the citing and cited articles, whereas the source of results and statements points to a more conceptual connection. Determining the precise value of these operational and conceptual relationships for citation evaluation requires further research, yet it highlights the importance of considering source meaning as a critical variable in assessing citation value.

Measurement of the citation function can also determine whether a citation serves an organic or perfunctory purpose. Studies have shown that the majority of citations are perfunctory (Chubin &Moitra 1975; Moravcsik & Murugesan 1975; Xu, Martin & Mahidadia 2014; Zhao and Strotmann 2020). Citations associated with utilization, explanation, and comparison functions generally indicate high citation value, as they suggest that the cited work is genuinely needed (organic citation). Citations that serve as related work represent a moderate need, positioned between organic and perfunctory, while those that serve an introductory function often yield lower citation values due to their perfunctory nature.

While these three meanings of citation sentences offer the potential for assessing citation value, existing literature has not yet developed a framework for such measurements. Further research is necessary to establish a citation-value assessment framework grounded in the meaning of citations within the text, thereby shifting the focus of citation evaluation from quantity to quality.

## Conclusion

The simultaneous classification of multiple citation meanings can be accomplished using a multi-task deep learning approach. However, the challenges lie in the imbalanced class distribution within the dataset and the vast number of hyperparameter combinations. In this study, three types of MTL models are developed, namely shared-trunk, cross-stitched, and shared-private, combining three different architectures: CNN, LSTM, and BiLSTM. An analysis of time efficiency and algorithmic complexity reveals no significant discrepancies among the models. Through performance evaluation, it is determined that the MTL shared-private-layer CNN model is the most effective for analyzing three different meanings. This model achieves an F1 score > 0.80 for all classification tasks and attains the highest scores for role source (0.87) and function (0.90). However, for sentiment classification, the score (0.81) is lower than that of the single-task CNN (0.85). The vast data imbalance within the sentiment task poses a challenge for the constructed MTL model.

This study is limited to journal data from Indonesia, consisting of nearly 10,000 sentences. To ensure broader applicability, the dataset should be expanded to include citation-context

sentences from various journals and languages. Furthermore, the model's development should continue by incorporating advances in NLP, such as pre-trained language models.

## References

Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623-2631). https://doi.org/10.1145/3292500.3330701

Baig, Y. M., Oesterling, A. X., Xin, R., Yu, H., Ghosal, A., Semenova, L. & Rudin, C. (2021, June). Multitask learning for citation purpose classification. In *Proceedings of the Second Workshop on Scholarly Document Processing* (pp. 134-139). Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.sdp-1.18.pdf

Budi, I. & Yaniasih, Y. (2023). Understanding the meanings of citations using sentiment, role, and citation function classifications. *Scientometrics*, 128(1), 735-759. https://doi.org/10.1007/s11192-022-04567-4

Chen, J., Qiu, X., Liu, P. & Huang, X. (2018). Meta multi-task learning for sequence modeling. *AAAI'18/IAAI'18/EAAI'18: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence* (pp. 5070 – 5077). New Orleans, Louisiana, USA: AAAI Press.

Cohan, A., Ammar, W., Van Zuylen, M. & Cady, F. (2019, June). Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers)* (pp. 3586-3596). Minneapolis, Minnesota. Association for Computational Linguistic https://doi.org/10.18653/v1/N19-1361

Crawshaw, M. (2020). Multi-task learning with deep neural networks: A survey. https://arxiv.org/abs/2009.09796

Ghosh, S., Das, D. & Chakraborty, T. (2018). Determining sentiment in citation text and analyzing its impact on the proposed ranking index. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9624 LNCS* (pp. 292–306). https://doi.org/10.1007/978-3-319-75487-1_23

Chubin, D. E. & Moitra, S. D. (1975). Content analysis of references: adjunct or alternative to citation counting? *Social Studies of Science*, 5(4), 423-441. https://doi.org/10.1177/030631277500500403

Iqbal, S., Hassan, S. U., Aljohani, N. R., Alelyani, S., Nawaz, R. & Bornmann, L. (2021). A decade of in-text citation analysis based on natural language processing and machine learning techniques: an overview of empirical studies. *Scientometrics*, 126(8), 6551-6599. https://doi.org/10.1007/s11192-021-04055-1

Justus, D., Brennan, J., Bonner, S. & McGough, A. S. (2018, December). Predicting the computational cost of deep learning models. In *2018 IEEE international conference on big data (Big Data)* (pp. 3873-3882). IEEE. https://doi.org/10.1109/BigData.2018.8622396

Lever, J., Krzywinski, M. & Altman, N. (2016). Classification evaluation. *Nature Methods*, 13(8), 603–605. https://doi.org/10.1038/nmeth.3945

Liu, P., Qiu, X. & Huang, X. J. (2017, July). Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (volume 1: long papers) (pp. 1-10). https://doi.org/10.18653/v1/P17-1001

Misra, I., Shrivastava, A., Gupta, A. & Hebert, M. (2016). Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3994-4003). https://doi.org/10.1109/CVPR.2016.433

Moravcsik, M. J. & Murugesan, P. (1975). Some Results on the Function and Quality of Citations. *Social Studies of Science*, 5(1), 86–92. https://doi.org/10.1177/030631277500500106

Nazir, S., Asif, M., Ahmad, S., Aljuaid, H., Iftikhar, R., Nawaz, Z., & Ghadi, Y. Y. (2022). Important citation identification by exploding the sentiment analysis and section-wise in-text citation weights. *IEEE Access*, 10, 87990-88000. https://doi.org/10.1109/ACCESS.2022.3199420

Nicholson, J. M., Mordaunt, M., Lopez, P., Uppala, A., Rosati, D., Rodrigues, N. P., Grabitz, P. & Rife, S. C. (2021). Scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quantitative Science Studies*, 2(3), 882–898. https://doi.org/10.1162/qss_a_00146

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. https://doi.org/10.48550/arXiv.1706.05098

Sewak, M., Sahay, S. K. & Rathore, H. (2020, November). Assessment of the relative importance of different hyper-parameters of LSTM for an IDS. In *2020 IEEE REGION 10 CONFERENCE (TENCON)* (pp. 414-419). IEEE. https://dx.doi.org/10.1109/TENCON50793.2020.9293731

Sharma, A., Van Rijn, J. N., Hutter, F. & Müller, A. (2019, October). Hyperparameter importance for image classification by residual neural networks. In *International conference on discovery science* (pp. 112-126). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-33778-0_10

Su, X., Prasad, A., Kan, M. & Sugiyama, K. (2019). Neural Multi-Task Learning for Citation Function and Provenance. *JCDL '19: Proceedings of the 18th Joint Conference on Digital Libraries* (pp. 394-395). https://doi.org/10.1109/JCDL.2019.00122

Swales, J. (1986). Citation analysis and discourse analysis. *Applied Linguistics*, 7(1), 39-56. https://doi.org/10.1093/applin/7.1.39

Tovedal, S. (2020). *On the Effectiveness of Multi-Task Learning: An evaluation of multi-task learning techniques in deep learning models*. Master's thesis. Umeå University, Sweden. Retrieved from http://www.diva-portal.org/smash/record.jsf?pid=diva2:1442398

Wang, X., Lyu, J., Dong, L. & Xu, K. (2019). Multitask learning for biomedical named entity recognition with cross-sharing structure. *BMC Bioinformatics*, 20(1), 427. https://doi.org/10.1186/s12859-019-3000-5

Xu, H., Martin, E. & Mahidadia, A. (2014). Contents and time sensitive document ranking of scientific literature. *Journal of Informetrics*, 8(3), 546–561. https://doi.org/10.1016/j.joi.2014.04.006

Yaniasih, Y. & Budi, I. (2021). Systematic Design and Evaluation of a Citation Function Classification Scheme in Indonesian Journals. *Publications*, 9(27), 27. https://doi.org/10.3390/publications9030027

Yousif, A., Niu, Z., Chambua, J. & Khan, Z. Y. (2019). Multi-task learning model based on recurrent convolutional neural networks for citation sentiment and purpose classification. *Neurocomputing*, 335(28), 195-205. https://doi.org/10.1016/j.neucom.2019.01.021

Zhang, Y. & Yang, Q. (2018). An overview of multi-task learning. *National Science Review*, 5(1), 30-43. https://doi.org/10.1093/nsr/nwx105

Zhang, Y., Zhao, R., Wang, Y., Chen, H. & Mahmood, A. (2022). Towards employing native information in citation function classification. *Scientometrics*, 127(11), 6557-6577. https://doi.org/10.1007/s11192-021-04242-0

Zhao, D. & Strotmann, A. (2020). Deep and narrow impact: Introducing location filtered citation counting. *Scientometrics*, 122(1), 503-517. https://doi.org/10.1007/s11192-019-03280-z

Zhao, H., Luo, Z., Feng, C. & Ye, Y. (2019, July). A context-based framework for resource citation classification in scientific literatures. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1041-1044). https://doi.org/10.1145/3331184.3331348

Zhao, H., Luo, Z., Feng, C., Zheng, A. & Liu, X. (2019, November). A context-based framework for modeling the role and function of on-line resource citations in scientific literature. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 5206-5215). https://doi.org/10.18653/v1/d19-1524