A Deep Learning Framework to Identify Interconnectivity of Citation Networks

MSc Research Project
MSc Artificial Intelligence

Taylou Maniganze
Student ID: 22162071

School of Computing
National College of Ireland

Supervisor: Prof. Paul Stynes

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Taylou Maniganze
Student ID:	22162071
Programme:	MSCAI1 Jan23 Artificial Intelligence **Year:** 2023
Module:	MSc Research in computing
Supervisor:	Prof. Paul Stynes
Submission Due Date:	27/05/2024
Project Title:	A Deep Learning Framework to Identify Interconnectivity of Citation Networks.
Word Count:	**4241 Page Count: 11**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Taylou Maniganze

Date: 27/05/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Deep Learning Framework to Identify Interconnectivity of Citation Networks

Taylou Maniganze
22162071

Abstract

Citation networks are graphs that describe citations within a collection of scholarly articles. Citation analysis can be used to identify the amount of interconnectivity such as the citation structure, word refreshes, influence, and quality of research. Identifying the related citation to an author’s research is a challenge due to the number of papers that exist, and the datasets that are available have a lot of papers which leads to datasets being very huge.

This research proposes a Deep Learning Framework to identify interconnectivity of citation networks. The proposed framework combines Transformer models and Interconnectivity. The transformer models are trained on three different datasets C2D-I, ACL-ARC, and SciCite. This research compares transformer models BERT_base, BERT_large, and DistilBERT to identify the model with the highest accuracy. This research is of interest to research students who require help in identifying the most relevant research papers for their research area.

Keywords: Transformer models, word embeddings, citation analysis, citation network, BERT.

1 Introduction

Citation analysis is the measurement of a paper or publication frequency or patterns by counting how many times a paper or publication was cited by other researchers. It uses a directed graph of citations to show the link between one document and another, it also shows the properties of the document. Citation analysis identifies the impact of a publication by showing which researcher cited it or based their paper on it as well as to identify seminal works.

A Transformer model is a type of neural network architecture that is based on the parallel multi-head attention mechanism (Vaswani et al., 2017). It has two parts encoder and decoder the encoder takes care of the input sequence, and the decoder takes care of the output sequence. It works through sequence-to-sequence learning and the way it does that it takes a sequence of tokens in the input sequence and predicts the next word in the output through encoder layers. The encoder generates encodings that define which part of the input sequence are relevant to each other and then passes the encodings onto the next layer. The decoder then takes all the encodings and uses the derived content to generate the output sequence. It is characterized by requiring less training time than previous recurrent architectures, such as long short-term memory(LSTM) (Vaswani et al., 2017) and GRU. Transformer-based models have been predominantly adopted for training large language models on large datasets, such as Wikipedia corpus and Common Crawl, due to the parallelized processing of the input sequence (Vaswani et al., 2017) Transformer models have achieved breakthrough results in natural language processing, but their application to citation networks has the potential to provide researchers with unique insights and answers.

The aim of this research is to investigate to what extent a Deep Learning framework can identify the interconnectivity of a citation network using Transformers. To address the research question, the following specific sets of research objectives were derived:

Design a deep learning framework architecture that includes state-of-the-art Transformers. The framework will combine the Transformer models and interconnectivity. Integrate innovative techniques to enhance the ability of networks to capture complex interconnectivity within citations.

Implement the deep learning framework to the extent that the model understands and represents the relationships within the citation data. The implementation will be based on ensuring that the model was trained effectively and hyperparameters were selected carefully.

Evaluate the performance of the deep learning framework which is basically assessing how accurate the deep learning framework will perform within the citation network.

The major contribution of this research is a novel Deep Learning framework that combines Transformer models for detecting interconnectivity on citation network analysis.

This paper discusses deep learning models used for interconnectivity and application to citation network analysis of the research, in section 2 previous work on citation network analysis and transformer models is reviewed. Section 3 describes the methodology for preprocessing citation analysis data and representing it as input for transformer models and detecting interconnectivity. Section 4 presents the design components and implementation of transformer models for the interconnectedness detection task. Section 5 reports on the evaluation and comparison of the proposed models with baseline methods and existing approaches on recurrent citation network analysis. Section 6 concludes the research and discusses future work.

2 Related Work

There are many reasons as to why scientific content should cite other works which is referred to as citation intent. Some of the reasons include finding citation networks, relevant papers, citation intent classification, context analysis, and article recommendation. The citation network plays a big role in finding the impact factor. A couple of methods have been proposed mostly based on references and citation intent information to predict the connection between research articles by assuming if paperA cites paperB or authorA cites authorB there is a connection between the two regardless of the topic, reasons, and importance.

The examination of how to map scientific papers using the contextual proximity of citations using a pre-trained BERT model was presented by (Roman et al., 2022). The research gives very detailed information, it uses a technique based on the entire content of the research article to find the reasons behind citations by looking at the labels of the edges of a citation network. (Roman et al., 2022) converted the text data and classified the text into three intent classes background, method, and result. The research also developed its own dataset which was one of the datasets that were used in this paper C2D-I (Intent) available upon request from the big dataset C2D with 53 million instances used a transformer model to get the words contextual representation. The results from the proposed method got an F1 score of 0.89. However, this work encountered a couple of limitations in the replication of the paper, the dataset with the most instances that was mainly used by (Roman et al., 2022) and was said to contain three intent classes was missing the three intent classes, and since the paper's methodology is relying heavily on that specific dataset it resulted to the complication and inconsistency in the replication of the paper.

Another challenge this study encountered was in the replication of the methodology as the GitHub code provided by (Roman et al., 2022) had a couple of bugs even with careful adherence, the results obtained are different from those in the State-of-the-Art. This means that the difference in the results obtained in this paper and those in the State-of-the-Art need an in-depth analysis of the techniques used, we are looking at several techniques such as preprocessing procedures, configuration of parameters, and word representation model. The main goal of this literature review is to contribute to an ongoing discussion on research replication and transparency in the limitations encountered.

Iqbal et al. (2023) conducted experiments using deep learning models with three word embedding techniques: FastText, word2vec, and BERT. Two datasets ACL-ARC and SciCite were used to evaluate the performance of these models, the results were then compared with several studies that used content-based features for classification. The experiments revealed that Convolutional Neural Network (CNN) model, together with FastText word embeddings, performed better than the model used in state-of-the-art experiments with 0.97 precision. CNN and fastText both combined achieved 0.89 accuracy, 0.97 precision, and 0.90 F1 score which is important for citation network analysis. However, there is a need for experiments on larger datasets.

Generation of a citation context-aware citation analysis embeddings in the work of (Ohagi and Aizawa, 2022) suggested a masked paper prediction task technique that predicts from three categories that are given by the citing paper and the citation context, the categories are the cited paper, citing paper and the citation context. The technique trains SciBERT, a variation of BERT, the prediction of the cited paper the related citation, and the citing paper is done with a masked paper prediction task, it gets the connection from citation networks. This approach is performed by implementing a novel loss function that considers the context of a citation and the reasons around the nodes in the citation network. The new loss function is based on a clear understanding of the structure of citation networks. Two datasets, FullTextPeer-Read and AASC were used to demonstrate the suggested technique. (Ohagi and Aizawa, 2022) showed how this approach outperformed the existing method hyperdoc2vec, which is a current approach for citation context-aware citation network embedding, using citation-recommendation and paper-classification tasks. SciBERT has better results on the FullTextPeerRead dataset and BERT has better results on the AASC dataset. SciBERT has a lower macro F1 of 0.48 score on the FullTextPeerRead dataset and a higher macro F1 of 0.74 score on the AASC dataset. BERT on FullTextPeerRead performed a lower macro score of 0.46 and a higher macro F1 of 0.73 on the AASC dataset for paper classification.

The architecture of the model that is presented by (Chen, Yin, and Qiu, 2009) uses initial vectors from pre-trained GloVe and EMLo embeddings within a text semantics encoder on the ACL Anthology network dataset. To get the bidirectional semantics of input sequence words, the encoder combines dot-product attentions with Bidirectional Gated Recurrent Units (BiGRUs). In contrast to previous methods, which preserve a paper's single vector representation to obtain backward updating data, the suggested text semantics encoder takes care of the semantics of every updated word in the title. The model presented by Chen, Yin, and Qiu (2009) provides a semantic weight parameter through the citation encoder. The parameter allows semantics from the title of the citing paper and citation context to be extracted independently, which results in a complex comprehension of the citing paper's content. After that, the model calculates the citation probabilities using the weighted title-context semantics, which provides a more contextually aware method of citation recommendation. In all the evaluation metrics, BACR which AllenNLP 2.4.0 performs better than baseline models with a 7% increase.

A framework that focuses on automatic node-level feature extraction and prediction in the setting of dynamic graphs is proposed by (Abbas et al. 2023). The framework ranks focus on temporal features, which makes it flexible for generalization to several network types. The model can contain explicit node-level properties, which increases its applicability across multiple applications. (Abbas et al. 2023) also introduced a novel cost function designed to address ranking issues in graphs using probabilistic regression methods to enable optimization and enhance the model's suitability for real-world applications. The findings highlight the model's performance metrics, in particular when evaluating a past time window of 10 months and predicting 10 months into the future. AUC-0.974, Kendal's rank correlation tau-0.455, Precision-0.643, Novelty-0.0456, Temporal novelty-0.375, and NDCG-0.949 were all achieved.

Citation intent, the reason or goal behind an author citing another author, is another activity that can be carried out with citation network analysis. Understanding the nature and purpose of citations in scholarly discourse can be aided by considering the citation's goal. (Roman et al., 2021) proposed a classification mechanism of clustering text as one of the techniques for determining citation intent. The study shows how according to the semantic and grammatical similarity, words are represented as vectors in a high-dimensional space using the word embedding technique. The text surrounding a citation was represented as a vector by (Roman et al., 2021) using word embedding was clustered in three categories (method, result, and background). The study used K-MEANS and HDBSCAN algorithms and word embedding techniques like GloVe, Infersent, and BERT. They used their approach to examine a dataset of 10 million citation contexts from the C2D dataset and discovered that BERT-KMEANS embedding outperformed other approaches significantly with an 81% F1 score. A labeled dataset for citation intent analysis was also supplied, which can be used in future studies.

Citation relevance, or the degree of interconnectedness between citing and cited publications, is one of the tasks that citation network analysis can do. Researchers can use citation relevancy to find the most relevant papers for their study topic. The study of (Shahid et al. 2021) shows to what extent an extension of the direct citation model using in-text citations can Identify citation relevancy. The direct citation model is a mechanism that determines the relevance of two papers based on the number of citations they receive. This strategy, however, ignores the context and purpose of the citations. The methods of (Shahid et al., 2021) recommended using the frequency of in-text citations as a metric for citation relevance. The way they approached it they reasoned that the more a work is cited in another paper, the more relevant it is. They also classified the association between citing and cited publications as weak, medium, and strong. They used their method to evaluate a CiteSeerX dataset of 5,000 research publications and the experiments discovered that citation relevancy is a quadratic polynomial function of in-text citation frequency. This study compared their strategy to others as well, such as content-based, bibliographic coupling, and metadata-based models, and demonstrated how their method got better results. Results show that the more in-text citations the stronger the relationship between citing and cited papers when in-text citation of cited papers was greater or equal to 5 times which occurred 77% of the time. When in-text citation of cited papers was less than five times there was a weak relationship between cited papers and citing papers and that occurred 87% of the time. Their proposed method produced a normalized Discount Cumulative Gain value of 89%.

A deep learning model that uses Long Short Memory (LSTM) was proposed by (Prasad, Kaur and Kan, 2018). The deep learning model takes reference texts and classifies them in different categories, among those categories are location, journal, author, and date. The deep learning model that was proposed works in sequence-to-sequence to label the categories then it takes the word embeddings as inputs. After the experiments on both English and multilingual data, they obtained a P value less than 0.01 in two domains humanities and computing.

Hassan et al. (2018) compared their proposed deep learning citation context model with traditional machine learning models that were proposed by other researchers. Their study showed that most traditional models such as Naïve Bayes, SVM, RF, and KNN performed fairly well in predicting important and unimportant citations. Their proposed model was an implementation of a Long Short Term Memory (LSTM) model which used an optimized Keras implementation with six layers. Although their model performed well, one of their limitations was the definitions of the 'important' and 'unimportant' text as defined by the standard dataset. Citation network analysis could possibly overcome this limitation since the interconnectedness of a paper could enhance or not enhance its citation importance in this context. Their approach, according to the authors, does not scale well, so it needs a significantly more powerful model such as BERT to be able to handle larger datasets.

Chanwoo et al. proposed a solution to citation recommendations using BERT on modified versions of state-of-the-art datasets. The model for recommendation was a non-parametric probability model like the proposed method for this research. They created a citation encoder for linking a prediction by using citation relationships between papers as input values. This research attempts to further study the creation of citation networks which could improve the ability of recommender systems to provide fully context-aware recommendations. The authors used BERT's CLS-PAIR model within the context encoder. This research uses BERT's small model. However, they added the GCN model for citation relationship between papers while BERT was used for the learning presentation.

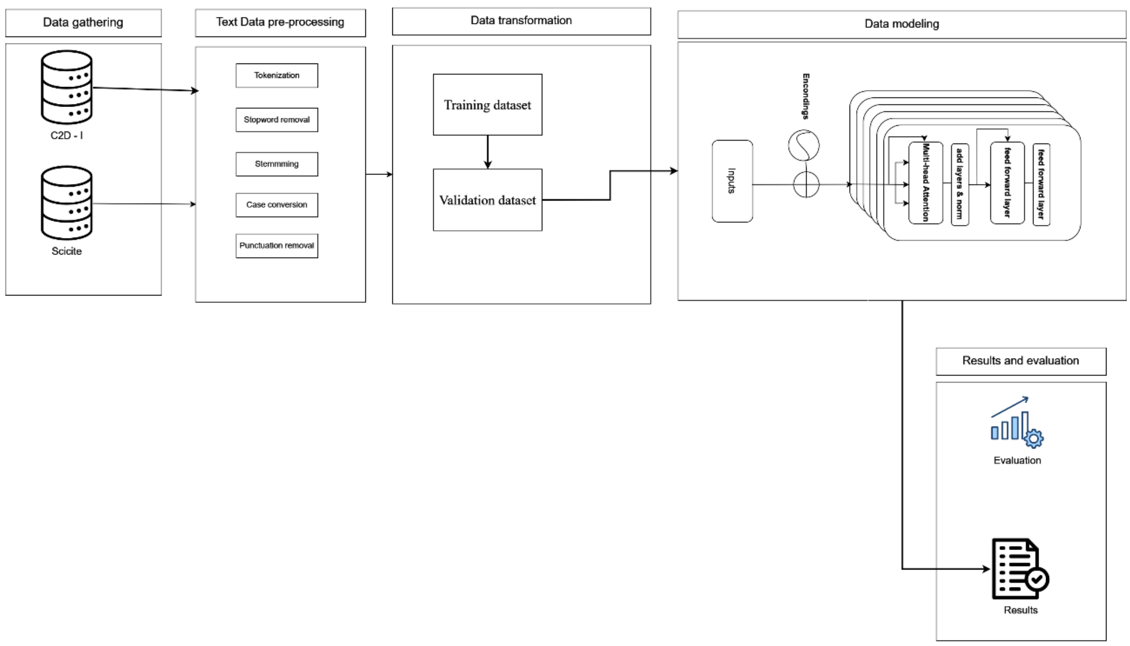
A systematic review, which is identifying, analyzing, and synthesizing the relevant literature on a specific research issue or topic, is another activity that can be performed using citation network analysis. A systematic review can help researchers summarize existing state-of-the-art knowledge and identify gaps and future research areas. The AI-based method presented by (Liu, Lo, and Kan, 2021) is one of the strategies proposed for systematic review with citation network analysis. The AI-based method automates systematic review procedures such as literature search, screening, extraction, analysis, and visualization through the use of natural language processing (NLP) and machine learning (ML) techniques. Furthermore, the AI-based strategy uses citation network analysis to discover important publications, significant authors, research clusters, and emerging trends in a topic. The AI-based technology can be used to examine citation networks of various areas and scales and objective literature review can be carried out.

Since 2012, neural networks have been good at vision tasks like identifying objects in a picture. However, for quite a while there was nothing good to analyze, translate, summarize, and generate text and this was a challenge since language is the primary way that humans use to communicate. Traditional models were never efficient in handling large volumes of text such as long paragraphs since they looked at text sequentially. They were not able to maintain the context of the entire text. Models such as RNN read through text sequentially so the main limitation was the ability to parallelize the training process. It was not very scalable even by adding more computational power such as GPUs. These limitations did not enable RNNS to be trained on large texts due to the exponential increase in training time. Transformers solve this issue with their scalable architecture and ability to parallelize the computations to handle extremely large text data.

3 Research Methodology

The research methodology consists of five steps namely text data gathering, text data pre-processing and tokenizing, data modeling, evaluation, and results as shown in Table 1

Table 1



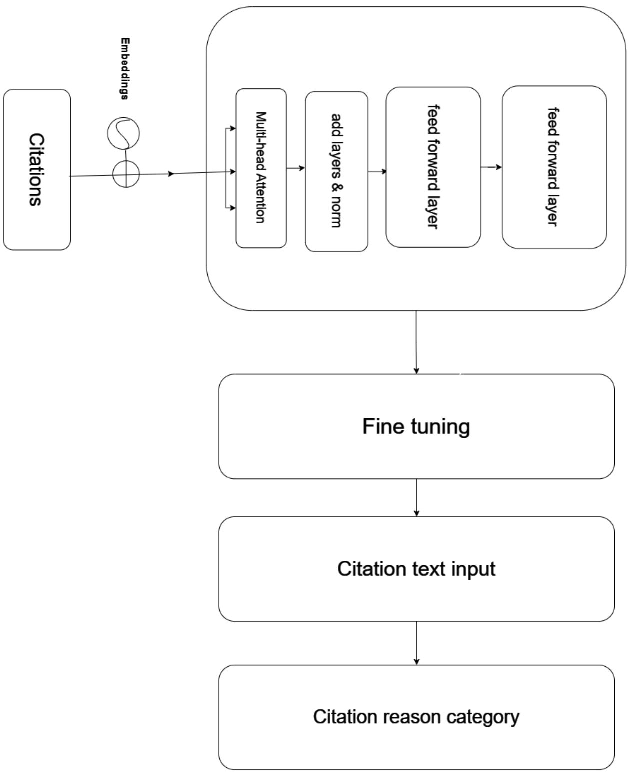
The first step, Data Gathering involves getting two text data datasets C2D-I (upon request)and Scicite.

The second step, Text data Pre-processing involves loading and cleaning the raw text, one dataset was checked manually as it had an unlabelled column. The techniques for pre-processing the text data involve stopwords and punctuation removal, stemming the tokens, and lowercasing. After that, the text data is encoded and tokenized using Bi-Directional Encoder Representation from Transformers BERT (Vaswani et al., 2017). The text is converted into input_ids which is suitable for the BERT model. The text that is encoded is then truncated to a max length of 512 tokens for the consistency of the input size for the model.

The third step, Text data modelling is based on the pre-trained BERT large model. The transformer model layer created from the TFAutoModel is used as the model's base. The sequence_output is extracted once the input_word_ids are sent to the transformer layer. The classification token (CLS) at the beginning of the sequence_output is chosen and regularized using a dropout layer. Lastly, the predicted probabilities for each class are obtained using a dense layer with a softmax activation function. After that splitting the encoded texts and related labels into training and testing datasets. The training dataset is then processed further by repeating, shuffling, and batching it.

The third step, Evaluation and Results involves evaluating the performance of the BERT model on each dataset. The test dataset is used for evaluation. Adam optimizer using the categorical cross-entropy loss function and accuracy are the metrics used for evaluation. Learning rate schedulers are implemented using the exponential decay function to adjust the learning rate during training. And 10 epochs are used.

4 Design Specification



The architecture is replicated on the Kaggle platform as the State-of-the-Art (Roman et al., 2022) the platform was selected for its strong hardware accelerator capabilities in deep learning tasks which are important in training heavy deep learning models. The same architecture was replicated on the AWS instance as well, as the replication was not giving better results. The BERT model requires a lot of memory, and the capability of TPUs and GPUs helps to improve the training process and increase model performance. When the model is given input features it should predict the citation reason of a cited publication. BERT uses the encoder and the decoder where the encoding task produces the embeddings of

the text input with intent information and the decoding task takes the embeddings from the encoding task and connects the input to the output.
BERT consists of two models, BERT large which contains 24 layers with 16 attention heads that are trained on 340 million parameters, and BERT base with 12 transformer blocks and 12 attention heads trained on 110 million parameters (Vaswani et al., 2017). This work trained BERT in two ways, first the important for most of the tasks is pre-training the second is fine-tuning for certain tasks.

The dataset that is mainly used in this paper is C2D-I with 30,000 because it contains more instances than the other two datasets SciCite and ACL-ARC. When the deep learning framework receives input texts it should predict the citation reasons such as background, method, and results of the referenced paper. Unlike RNN the deep learning transformer uses encoder-decoder architecture to pass input text one after the other. All words in a sentence are processed in parallel and capture the context of a sentence from right to left and left to right.

The hugging face (huggingface.co, 2024) provides the Transformers library which is used for working with state-of-the-art pre-trained transformer models including BERT. The hugging face's BERT is used for tokenization as well. The deep learning framework uses TensorFlow as the main deep learning framework for model development.

5 Implementation

The deep learning framework was implemented on Kaggle and AWS 12xlarge instance. All the libraries and packages were installed on the instance making sure that they are compatible. The libraries installed include Python, Tensorflow, TensorRT, Transformers and NLTK. First, the text data is converted into lowercase then applied stopwords and stemering. BERT Tokenizer is then initialized for tokenizing the input text data and preparing it for processing, tokenizing, and encoding the text data with BERT tokenizer is defined by the tokenize_and_encode function. It adds tokens, and pads, and the input data is returned. The encoded input_ids and labels are stored in X_ids and y after they have been tokenized and their maximum length has been set to first 128 then 256 and the recommended max length is 512. The training optimizer is Adam optimizer as it works well in most scenarios. The X_ids and y were split with a test size of 0.3 and a random state of 2020. For the deep learning framework testing purpose, the model is given a text containing references from a published paper and then tells which category the text belongs to, background, method, and results.

The deep learning model used some techniques to improve the results including a learning rate scheduler with a decay of 0.01 and then 0.1 and increased the number of epochs from 10 to 30 epochs.

6 Evaluation

The aim of this experiment is to replicate the experiments in the state-of-the-art on two datasets C2D-I dataset with 30,000 instances and the SciCite dataset with 11,020 instances with different transformer models. However, due to the incomplete Github code provided by (Roman et al., 2022), there are a lot of modifications that were made during experiments. The framework is trained on both datasets using Kaggle and AWS the figures below show the comparison of the results from AWS GPUs which performed better compared to Kaggle TPUs.

6.1 Experiment EC2

The results in Table (2) show the comparison of the models on GPUs. BERT_base model on the C2D-I dataset predicted the target labels 80% of the training which means it learned the patterns and features in the training data. the accuracy of the model is 50% on the validation data. BERT_large did not give good results and mostly stopped due to memory. However, the results of BERT_large on sciciteDS datasets, although there is overfitting performs well on the data that it was trained on. It learned patterns and features in training data with 98% accuracy and a validation accuracy of 79%.

Table 2: EC2 instance

	C2D-I			SciciteDS		
	BERT_base	BERT_large	DistilBERT	BERT_base	BERT_large	DistilBERT
Accuracy	0.80	-	0.50	0.98	0.98	0.98
Validation accuracy	0.48	-	0.39	0.80	0.80	0.80

6.2 Experiment on Kaggle TPUs

Table (3) shows the comparison of all the models used on TPUs and their performances. The results show that the models did not perform well on all the datasets, but because sciciteDS has only 8243 instances it improved a little bit after doing some fine-tuning. C2D-I did not improve after fine-tuning, and it remained with the same results.

Table 3: Kaggle TPUs

	C2D-I			SciciteDS		
	BERT_base	BERT_large	DistilBERT	BERT_base	BERT_large	DistilBERT
Accuracy	0.33	0.33	0.40	0.59	0.59	0.59
Validation accuracy	0.20	0.20	0.30	0.40	0.40	0.40

7 Conclusion and Future Work

The aim of this research was to design a deep learning framework that identifies interconnectivity within cited publications. This research provides a critical analysis of the State-of-the-art and discusses their limitations and identifying interconnectivity behind cited publications. Results demonstrate that BERT and DistilBERT on both datasets C2D-I and sciciteDS BERT performed better on GPUs compared to TPUs. BERT_base shows promise for accuracy if there is more work to be done on the model. A limitation of this study is limited access to powerful accelerators as transformer models require a huge amount of computational time for training.

This research can help researchers in identifying the most relevant research papers for their research area. This work can be improved by optimizing the models and applying more fine-tuning techniques to select a final model for the framework based on accuracy, loss, and size. Furthermore, extensive research can be carried out on this work to find the reasons behind cited papers based on other citation intent classes. More research must be carried out for better accuracy in the models and to build a fully working deep learning framework to identify interconnectivity.

References

Abbas, K. *et al.* (2023) 'Predicting the Future Popularity of Academic Publications Using Deep Learning by Considering It as Temporal Citation Networks', *IEEE Access*, 11, pp. 83052–83068. Available at: https://doi.org/10.1109/ACCESS.2023.3290906.

google-bert (BERT community) (2024). Available at: https://huggingface.co/google-bert (Accessed: 25 April 2024).

Hassan, S.-U. *et al.* (2018) 'Deep context of citations using machine-learning models in scholarly full-text articles', *Scientometrics*, 117(3), pp. 1645–1662. Available at: https://doi.org/10.1007/s11192-018-2944-y.

Iqbal, A. *et al.* (2023) 'Exploiting Contextual Word Embedding for Identification of Important Citations: Incorporating Section-Wise Citation Counts and Metadata Features', *IEEE Access*, 11, pp. 114044–114060. Available at: https://doi.org/10.1109/ACCESS.2023.3320038.

Jeong, C. *et al.* (2020) 'A context-aware citation recommendation model with BERT and graph convolutional networks', *Scientometrics*, 124(3), pp. 1907–1922. Available at: https://doi.org/10.1007/s11192-020-03561-y.

Kreutz, C.K. and Schenkel, R. (2022) 'Scientific paper recommendation systems: a literature review of recent publications', *International Journal on Digital Libraries*, 23(4), pp. 335–369. Available at: https://doi.org/10.1007/s00799-022-00339-w.

Ohagi, M. and Aizawa, A. (2022) 'Pre-trained transformer-based citation context-aware citation network embeddings', in *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*. New York, NY, USA: Association for Computing Machinery (JCDL '22), pp. 1–5. Available at: https://doi.org/10.1145/3529372.3533290.

Peters, M.E. *et al.* (2018) 'Deep contextualized word representations'. arXiv. Available at: http://arxiv.org/abs/1802.05365 (Accessed: 15 April 2024).

Prasad, A., Kaur, M. and Kan, M.-Y. (2018) 'Neural ParsCit: a deep learning-based reference string parser', *International Journal on Digital Libraries*, 19(4), pp. 323–337. Available at: https://doi.org/10.1007/s00799-018-0242-1.

Shahid, A. *et al.* (2020) ‘Insights into relevant knowledge extraction techniques: a comprehensive review’, *Journal of Supercomputing*, 76(3), pp. 1695–1733. Available at: https://doi.org/10.1007/s11227-019-03009-y.
Shahid, A. *et al.* (2021) ‘Extension of Direct Citation Model Using In-Text Citations’, *Computers, Materials & Continua*, 66(3), pp. 3121–3138. Available at: https://doi.org/10.32604/cmc.2021.013809.
Vaswani, A. *et al.* (2023) ‘Attention Is All You Need’. arXiv. Available at: http://arxiv.org/abs/1706.03762 (Accessed: 23 April 2024).