Intermediate Update
SciCite

Guo Qingrui
Liu Yuyao
Pandya Poorva Harshang
Shen Shuyuan
Wang Yuehan
{e0950138, e0950438, e0559832, e0950159, e0950407}@u.nus.edu

Project Mentor: Hai Ye <hye.me@outlook.com>

V4.0 (202303).

2. Abstract

Citations are crucial in scientific works as they help to position a new publication. The goal of Scicite project is to classify the intent of a citation in scientific paper.

After the proposal of our project, we’ve done sufficient research about citation classification and understood the dataset. Furthermore, we have formed an overall graph of how we can conduct the experiment by reading the paper [Cohan et.al 2019] carefully, and learning how to conduct a better ablation study in order to design a model.

3. Motivation

- Background of publications citation:

  The report of the National Science Foundation in the end of 2021 shown that the number of scientific publications had grown constantly over past 20 years, from around 1.0 million articles to 2.8 million articles, which means citations play a crucial role in understanding the lineage and evolution of a field.

- Worthiness of this topic:

  Intent classification of a citation is critical for machine reading of individual publications and automated analysis of the scientific literature.

3. Motivation

- Real world example:

  When you go through a paper and want to explore any further improvement after the paper, citation intention classification can help us figure out whether any research paper has only used a particular section of this paper, or has it made further research on the model. This can also save workload and time for scientific workers.

- Overall, since citation plays an important role to analyze the impact of a scientific research, classifying the intent of a citation might be useful to improve automated analysis. Furthermore, by classifying the sentiment of the intent will help analyze the impact of a academic literature.

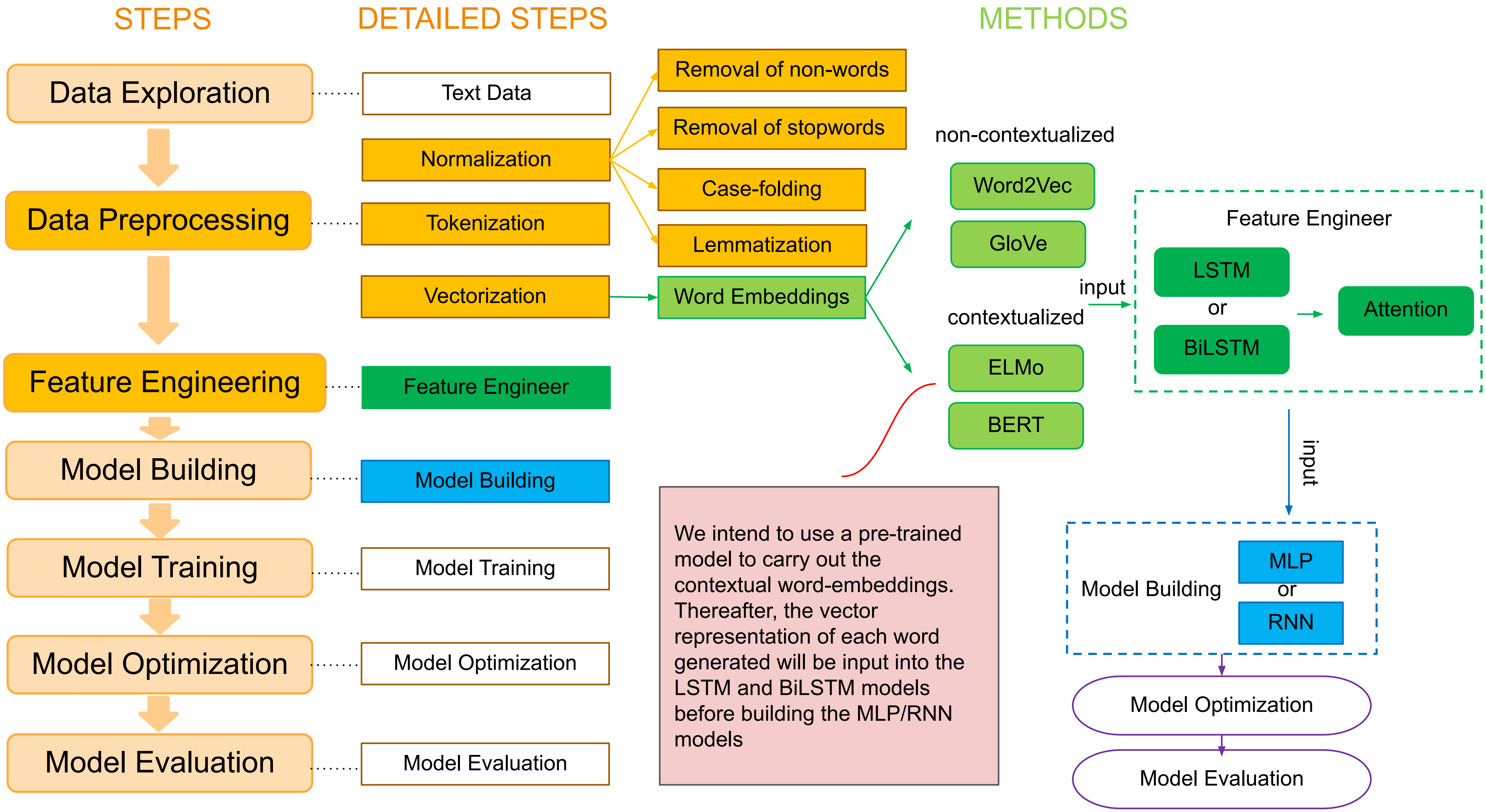
4. Task Statement

We are using the SciCite dataset, which includes 10,104 instances (8243 instances for the training dataset and 1861 instances for the testing). The main task is to do text classification (i.e., citation classification). The details of the dataset are in the ‘Resources’ section.
For this project, basically consider ‘string’ as the input which indicate the sentence that refers to the citation and ‘label’ as the output which is the classification.

String	label
For women with competing care-giving responsibilities, this involves allowing them to be accompanied by their children, providing on-site childcare or offering childcare subsidies [64,68,79].	background
We used an active contour algorithm [10] to segment the organs from 340 coronal slices over the two patients.	method
Consistent with previous reports (al-Khodairy et al., 1995; Seufert et al., 1995; Tanaka et al., 1999; Biggins et al., 2001), our data showed that Smt3 and Ubc9 have pivotal functions during mitosis	result

5. Proposed Method

We design the whole process of the project, including 7 main steps, data exploration, data preprocessing, feature engineering, model building, model training, model optimization and evaluation.
Detailed description of structure is as below.



STEPS

DETAILED STEPS

METHODS

Data Exploration

Data Preprocessing

Feature Engineering

Model Building

Model Training

Model Optimization

Model Evaluation

Text Data

Normalization

Tokenization

Vectorization

Feature Engineer

Model Building

Model Training

Model Optimization

Model Evaluation

Removal of non-words

Removal of stopwords

Case-folding

Lemmatization

Word Embeddings

non-contextualized

Word2Vec

GloVe

contextualized

ELMo

BERT

input

Feature Engineer

LSTM

or

BiLSTM

Attention

input

Model Building

MLP

or

RNN

Model Optimization

Model Evaluation

We intend to use a pre-trained model to carry out the contextual word-embeddings. Thereafter, the vector representation of each word generated will be input into the LSTM and BiLSTM models before building the MLP/RNN models

5. Proposed Method — Non-contextualized Embeddings

Word2Vec

The basic idea behind word2vec is to represent each word as a high-dimensional vector in a continuous vector space, where similar words are located closer to each other than dissimilar words. The word vectors are learned by training a neural network on a large corpus of text, where the network tries to predict the probability of a word given its surrounding context.

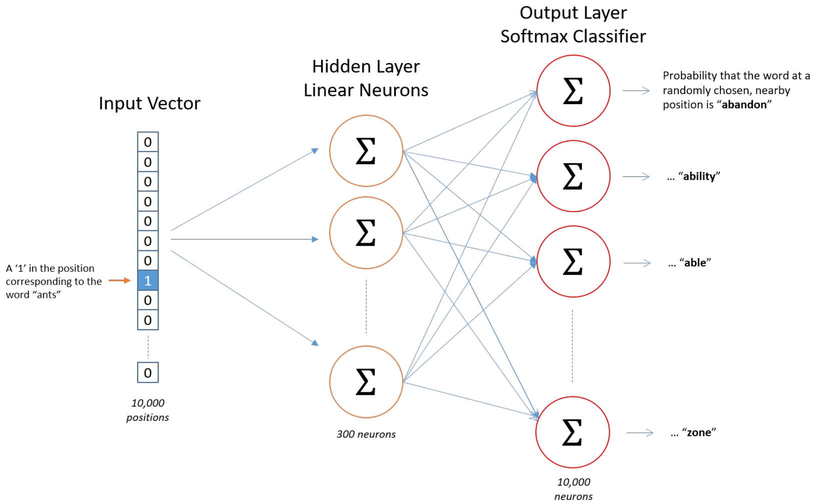


Image taken from http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Glove

The basic idea of GloVe is to learn the vectors by essentially doing some sort of dimensionality reduction on the co-occurrence counts matrix. It start by constructing a matrix with counts of word co-occurrence information, each row tells how often does a word occur with every other word in some defined context-size in a large corpus. This matrix is then factorize, resulting in a lower dimension matrix, where each row is some vector representation for each word.

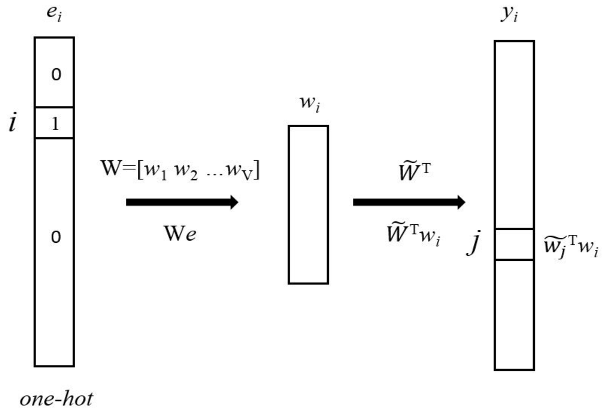


Image taken from
https://www.researchgate.net/figure/The-model-architecture-of-GloVe-The-input-is-a-one-hot-r

5. Proposed Method — Contextualized Embeddings

ELMo

The main idea of the ELMo can be divided into two main tasks, first it train an bi-directional LSTM network language model on some corpus, and then use the hidden states of the LSTM for each token to generate a vector representation of each word. The main advantage of ELMo is that it generates context-sensitive word embeddings that capture the syntactic and semantic nuances of words in their context.

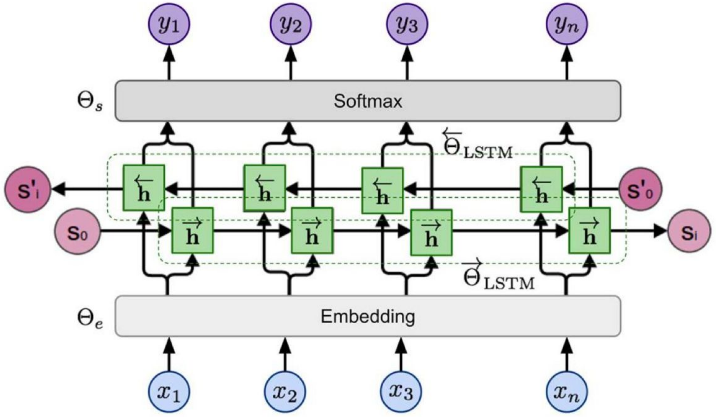


Image taken from https://medium.com/@kashyapkathrani/all-about-embeddings-829c8ff0bf5b

Bert

The basic idea behind BERT is to pre-train a large-scale transformer-based language model on a large corpus of text data to generate contextualized word embeddings that can be fine-tuned on specific downstream natural language processing tasks. By considering the entire context of a word in a sentence, BERT can capture complex linguistic phenomena and improve the accuracy of NLP applications.

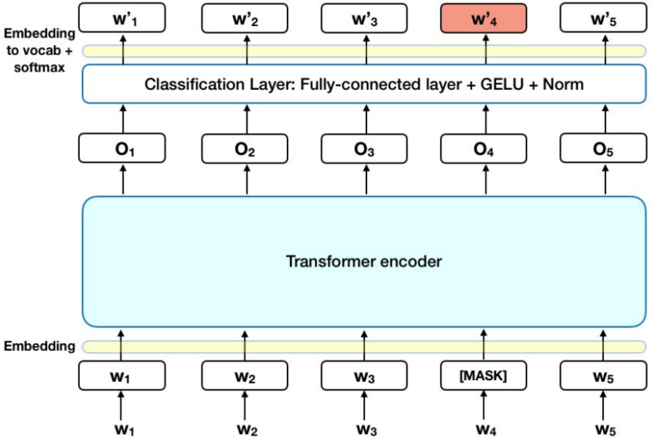


Image taken from https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

5. Proposed Method — Neural Networks

LSTM

The basic idea of LSTM is a type of recurrent neural network that addresses the vanishing gradient problem in vanilla RNNs through additional cells, input and output gates. Intuitively, vanishing gradients are solved through additional additive components, and forget gate activations, that allow the gradients to flow through the network without vanishing as quickly.

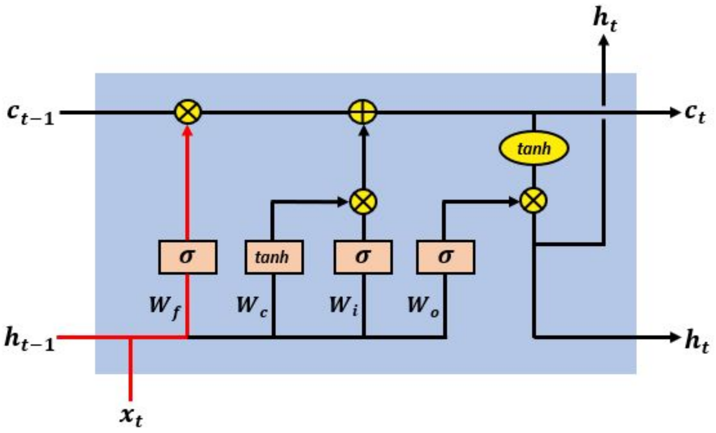


Image taken from https://paperswithcode.com/method/lstm

BiLSTM

The basic idea of BiLSTM, is a sequence processing model that consists of two LSTMs: one taking the input in a forward direction, and the other in a backwards direction. BiLSTMs effectively increase the amount of information available to the network, improving the context available to the algorithm.

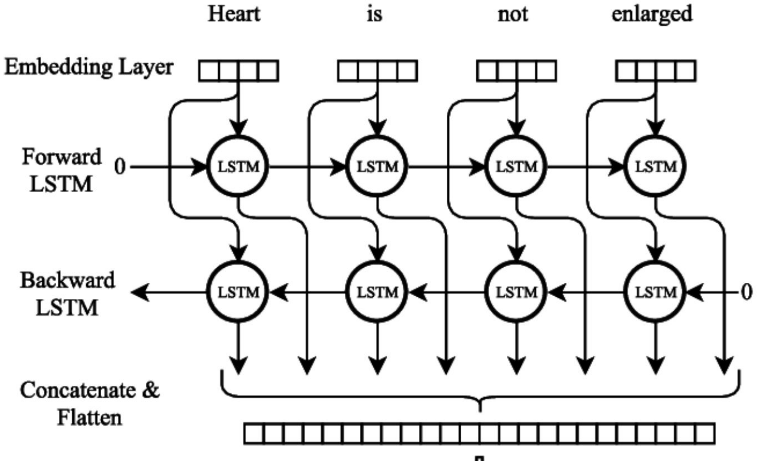


Image taken from https://paperswithcode.com/method/bilstm

6. Progress — Overview

We’ve completed data exploration and preprocessing. Our team members studied the target paper and methods that are often used when it comes to text classification. Thus, we can form a clear schedule (path) or how to conduct the project.

1. Research: Learn the methods proposed by other paper, how did other papers conduct their project and to further build our understanding of how to improve the model performance.
2. Data Exploration: Have a clear understanding of the dataset we are going to deal with.
3. Data preprocessing: Cleanse the data for a better training performance.

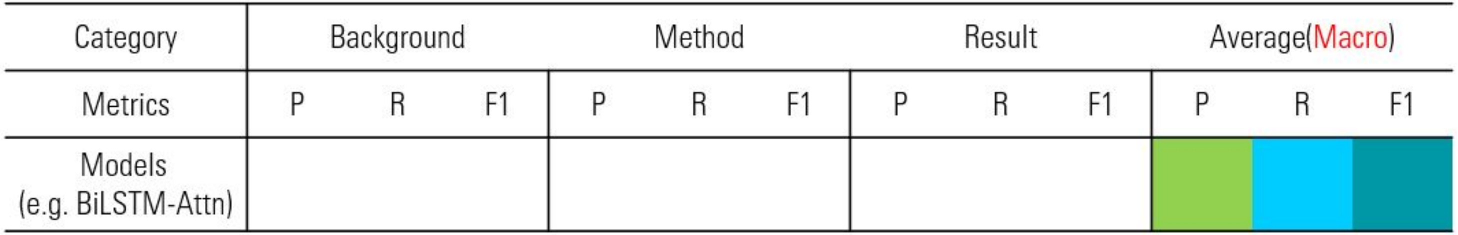
6. Progress — Detailed Task

	Task	Members	Status
1	Learn basic Python library to support further researching, reproducing and extension	All	Finished
2	Form a general scene of Text Classification tasks by researching on papers with high citations; Familiarize some milestone sequence model structures(LSTM, Transformer, BERT, etc.);	All	Finished
3	Data Exploration	All	Finished
4	Discussion and investigation of the methodology (i.e., feature engineering, model, evaluation metric) that can be used	All	Finished
5	Set up the possible ablation study	All	Finished
6	Data Normalization	All	In Progress

7. Proposed Evaluation

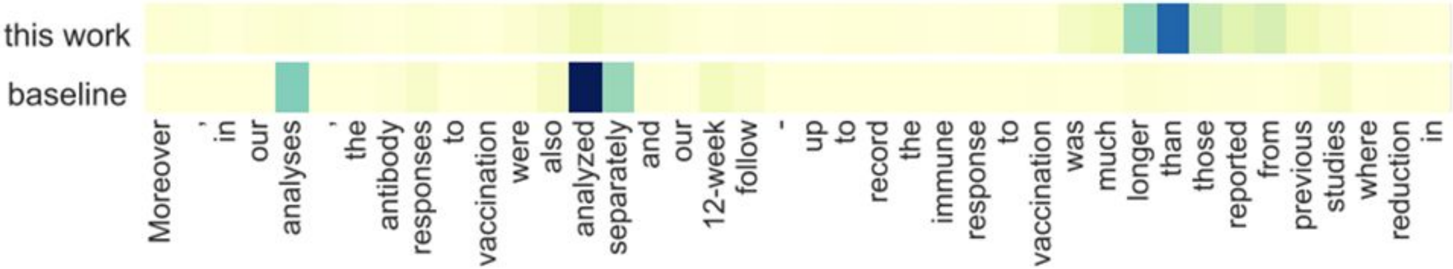
After investigating other similar tasks, we summarize some plausible evaluation metrics to measure as well as illustrate prediction performance:

- Precision, Recall, F1 (widely-used, usually take macro F1 as the final one)



Category	Background			Method			Result			Average(Macro)		
Metrics	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Models (e.g. BiLSTM-Attn)										🟩	🟦	🟦

- Confusion Matrices(Optional)
- Visualized Attention Weights(for attn methods)



Pictures: https://arxiv.org/pdf/1904.01608v2.pdf

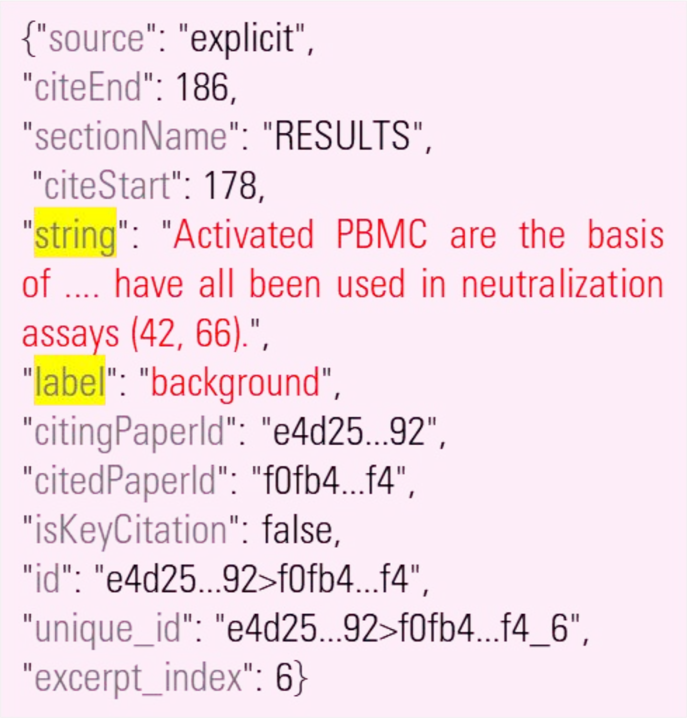
8. Resources - Dataset: SciCite

Only SciCite will be used during this project. (See: https://github.com/allenai/scicite)

- ➔ SciCite is a labeled citation texts dataset for Citation Intent Classification tasks
- ➔ Introduced by: *Cohan et al. in Structural Scaffolds for Citation Intent Classification in Scientific Publications* (https://arxiv.org/pdf/1904.01608v2.pdf)
- ➔ Features: more **coarse-grained**(reduce to 3 classes) and **general-domain** compared with existing datasets
- ➔ Format: JSONL(each line is a JSON for one citation)
- ➔ Labels(distribution): Background(0.58), Method(0.29), ResultComparision (0.13), indicating the purpose of each citation
- ➔ #Instances = 11,020 or 10,104; originated from 6,627 papers in Computer Science and Medicine domains

8. Resources - Dataset: SciCite

➔ Description & Sample:



```
{"source": "explicit",
"citeEnd": 186,
"sectionName": "RESULTS",
 "citeStart": 178,
"string": "Activated PBMC are the basis
of .... have all been used in neutralization
assays (42, 66).",
"label": "background",
"citingPaperId": "e4d25...92",
"citedPaperId": "f0fb4...f4",
"isKeyCitation": false,
"id": "e4d25...92>f0fb4...f4",
"unique_id": "e4d25...92>f0fb4...f4_6",
"excerpt_index": 6}
```

Key Name	Description
string	The string of text associated with the citation. (String)
sectionName	The name of the section the citation is found in. (String)
label	The label associated with the citation. (String)
isKeyCitation	A boolean value indicating whether the citation is a key citation. (Boolean)
~~**label2**~~	~~The second label associated with the citation. (String)~~
citeEnd	The end index of the citation in the text. (Integer)
citeStart	The start index of the citation in the text. (Integer)
source	The source of the citation. (String)

8. Resources - Dataset: Software/Hardware Issues

1.Compute

➔ Hopefully, we can make use of NUS computational resources to train Scicite scaffold models (otherwise can try Colab for free or local CPU)

2.Software/Libraries

➔ **AllenNLP**: To reproduce multitask model in SciCite paper, a recommended library is AllenNLP, an open-source NLP research library built on **PyTorch**

➔ **Pytorch**: To support baseline and extension methods like BERT, (Bi)LSTM.

3.Plan and Guidance

➔ For implementation, we will attempt our best to write codes by ourselves.

➔ Testing citations and a readme file will be included to help others run for replicability.

9. Schedule / Role Assignment

TimeLine	Main Task	Poorva	Yuyao	Qingrui	Yuehan	Shuyuan
Week 4-5 (1.30-2.12)	Previous Investigation	Project teams assembled; Choose SciCite dataset and the corresponding NLP task; Learn basic ML-related operations and functions under framework such as PyTorch to equip ourselves ML techniques and coding abilities;				
Week 6 (2.13-2.19)	Previous Investigation	Form a general scene of Text Classification tasks by researching on papers with high citations; Familiarize some milestone sequence model structures (LSTM,Transformer, BERT, etc.);Stipulate a long-term plans including different stages and results we have to achieve, the overall structure of this research project to fulfill; Assign work to each group member				
Week Recess (2.20-2.26)	Previous Investigation	Carefully research on the SciCite paper; Get explicit theoretical understandings of the micro model architecture on the main task		Conduct background research on pre-trained models (Word-Embedding)	Find extension method based on SciCtie(SciBERT, etc.) and alternative methods(DocBERT, AttentionCNN etc) which can apply to the same task	
Week 7 (2.27-3.04)	Data Preprocessing	Conduct data exploration and data preprocessing part; Complete Group intermediate update presentation				

9. Schedule / Role Assignment

TimeLine	Main Task	Poorva	Yuyao	Qingrui	Yuehan	Shuyuan
Week 8-9 (3.05-3.19)	Feature Engineering Model Building, Training and Optimizing	Implement Word2Vec	Implement ELMo	Implement GloVe	Implement BERT	Implement feature extraction methods(LSTM,BiLSTM)
		Inner-group discussion about research progress; Update experiment results and findings				
Week 10 (3.20-3.26)		Keep the framework but revise micro-structures in experiments		Implement alternative methods to make comparison to previous models;		
Week11-12 (3.27-4.6)	Model Evaluation and report	Complete experiment (implementation) part of the report: Explain the model architecture of proposed methods; Summarize experiment settings;		Clarify background and task, describe dataset, organize baseline	Try to explain model by visualizing trained parameters and illustrating structures; Draw conclusions and insights of the overall project	

10. Acknowledgements

Paper with Code: https://paperswithcode.com/dataset/scicite
Github Page of Scicite Paper: https://github.com/allenai/scicite

References

[1] Cohan, A., Ammar, W., Van Zuylen, M. and Cady, F., 2019. Structural scaffolds for citation intent classification in scientific publications. arXiv preprint arXiv:1904.01608.

[2] Mercier, D., Rizvi, S.T.R., Rajashekar, V., Dengel, A. and Ahmed, S., 2020. ImpactCite: an XLNet-based method for citation impact analysis. arXiv preprint arXiv:2005.06611.

[3] ROMAN, Muhammad, et al. Citation intent classification using word embedding. Ieee Access, 2021, 9: 9982-9995.

[4] MOTRICHENKO, Dmitry; NEDUMOV, Yaroslav; SKORNIAKOV, Kirill. Bag of Tricks for Citation Intent Classification via SciBERT. In: 2021 Ivannikov Ispras Open Conference (ISPRAS). IEEE, 2021. p. 120-126.

[5] NICHOLSON, Josh M., et al. Scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. Quantitative Science Studies, 2021, 2.3: 882-898.

[6]Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[7]Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

References

[8] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. North American Chapter of the Association for Computational Linguistics.
[9]Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[10] BERREBBI, Dan; HUYNH, Nicolas; BALALAU, Oana. GraphCite: Citation Intent Classification in Scientific Publications via Graph Embeddings. In: Companion Proceedings of the Web Conference 2022. 2022. p. 779-783.

Version History

V1.0 Feb 10, 2023 – Initial version, basic structure of content

V2.0 Feb 17, 2023 – Second version, Motivation, Task Statement and Schedule part.

V3.0 Feb 25, 2023 – Third version, Methodology part

V4.0 Mar 2, 2023 – Final version, Progress and Resources part