Amrita_CEN_NLP@SDP2021 Task A and B

Isha Indhu S, Kavya S Kumar, Lakshaya Karthikeyan, Premjith B, Soman K.P
Center for Computational Engineering and Networking (CEN)
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
b_premjith@cb.amrita.edu

Abstract

The purpose and influence of a citation are important in understanding the quality of a publication. The 3c citation context classification shared task at the Second Workshop on Scholarly Document Processing aims at addressing this problem. This paper is the submission of the team Amrita_CEN_NLP to the shared task. We employed Bi-directional Long Short Term Memory (Bi-LSTM) networks and a Random Forest classifier with fasttext embedding for modelling the aforementioned problems by considering the class imbalance problem in the data.

1 Introduction

In the evolution of the Information Age, where colossal amounts of research papers and scientific literature are now available, the need for a method which measures the scientific impact of a paper has become paramount. One such method is citation analysis. Citations are defined as a reference to the source of information used in one's research. The conventional approach to citation analysis involves utilising the frequency of citations Zhou et al. (2020) while treating all citations equally. This methodology provides a vague or even inaccurate overview of scientific development.

There are restrictions on the type of conclusions that can be drawn from citation counts, as many of the intricacies of citations are connected with the quality of the paper cited, and the context in which the citation is made. Two essential features for gauging the quality of a research paper are discerning the intent of citation and its level of influence. These features are the crux of the 3C shared task which has been split into two subtasks – Purpose (A) and Influence (B). Purpose is a multi-class classification problem which classifies the cited title based on how it is related to the citing paper. Meanwhile, influence is a binary-class classification problem which classifies whether the cited title is merely incidental or actually influential to the citing paper.

This paper reports the submissions of the Amrita_CEN_NLP team for the 3C Citation Context Classification shared task Kunnath et al. (2021). We used deep learning and machine learning models developed using Bi-LSTM and Random Forest algorithms Liaw et al. (2002), Premjith et al. (2019a), Premjith et al. (2019b) to complete the subtasks. They will be elaborated upon in Section 4.

2 Literature Review

Though the importance of categorizing scientific literature according to context is apparent, the reported amount of research that has been carried out is insufficient. Along with a classification model, Teufel et al. (2006) also proposed an annotation scheme for the categorization of the citations. 12 classes were considered for annotation. From 116 articles, 2829 citation samples were gathered. These were used to train the machine learning model. 113K algorithms were used for classification with hand-engineered features. One of such features was cue phrases. Features such as pattern-based features, topic-based features, and prototypical argument features were used by D. Jurgens Jurgens et al. (2018) to separate the documents into its 6 corresponding classes. The RandomForest algorithm was used for classification. Cohan et al. (2019) also utilised Glove, ELMO word embedding features, and Bi-LSTM with attention models to aid in the classification of the citations. Kunnath et al. (2020) organized the first shared task on citation classification in 2020, where different teams came up with different approaches to solve 3c classification problem.

3 Dataset Description

The training data that has been provided in Kaggle as a part of 3C shared task contains 3000 instances

annotated using the ACT platform, whereas the test set contains 1000 datapoints. The dataset is in csv format and contains the following fields: Unique Identifier, COREID of Citing Paper (Name of the text files in the full text dump), Citing Paper Title (Research Paper in consideration), Citing Paper Author (Author of the Citing Paper), Cited Paper Title (Paper cited by the Citing Paper), Cited Paper Author (Author of the Cited Paper), Citation Context (Citation Context represents the sentence containing the citations), and the Class Labels. The training set for Subtask-A is highly imbalanced, where the class BACKGROUND contains 54.93% of the total data, and the class FUTURE comprises a mere 2.07% of the data. The training data for the Subtask-B is balanced and has two class labels. Labels, description and the percentage of share in the dataset for subtask-A and subtask-B are shown in Tables 1 and 2, respectively.

4 System Description

This section describes the approaches used by Amrita_CEN_NLP to implement the models for both subtask-A and B. We experimented with machine learning as well as deep learning algorithms to complete the tasks. The workflow of the approach is given below.

1. Preprocessing
2. Feature extraction and Classification
3. Result analysis

4.1 Preprocessing

This step involves the data cleaning part. We considered both "cited title" and "citation context" for the analysis. After lowercasing all the characters in the input text, stop words were removed. At the same time, we didn't remove stop words for the deep learning models. This step was followed by deleting all the characters other than alphanumeric symbols from the text.

4.2 Feature Extraction and Classification

We employed both Random Forest and Bi-LSTM models for completing this task.

4.2.1 Random Forest Model

The Random Forest classifier used a Term Frequency-Inverse Document Frequency (TF-IDF) Chacko et al. (2019) with Singular Value Decomposition (SVD) Reshma et al. (2018) features for classification, as well as the fasttext embedding Bojanowski et al. (2017). For computing the TF-IDF values, we considered the unigram probabilities. SVD was applied to reduce the dimensionality of the feature vector by considering the first 50 components. The fasttext model was trained over the available training and testing data to generate the embeddings of dimension 300. A cost-sensitive learning approach Premjith and Soman (2020) was utilized to deal with the class imbalance problem.

4.2.2 Bi-LSTM Model

This work also used a Bi-LSTM for the classification. The workflow is as follows

1. Tokenize the input text into words
2. Use <OOV> token to handle any Out-of-Vocabulary words in the test data
3. Converted the words into integer indices by keeping all the words in the corpus for the analysis
4. Appended zeros at the end of each integer sequence to make the sequence length equal
5. Labels were converted into a one-hot encoded representation
6. Used Bi-LSTM with softmax layer for the classification of the text into different categories.

The hyperparameters used for building the model is explained in Table 3.

4.3 Result Analysis

Tables 4 and 5 show the performance of the models submitted for subtask-A and subtask-B, respectively. In the Tables Bi-LSTM, D and CW stand for Bi-LSTM, Dropout and Class weight, whereas 128 and 64 represent the number of hidden layer neurons in the Bi-LSTM. For the subtask-A, the Bi-LSTM + D + CW + 128 model outperformed other models both in private and public scores. In subtask-B, even though random forest classifier with fasttext embedding exhibited the highest private score, it is Bi-LSTM + D + CW + 64 scored the maximum in public score.

5 Conclusion

This working note describes the submission of Amrita_CEN_NLP at the 3c citation context classification task at Second Workshop on Scholarly Document Processing. Our team participated in both

Class Label	Description	Percentage share
Incidental	The cited title is incidental to the citing paper	52.27
Influential	The cited title is influential to the citing paper	47.73

Table 1: Influence Class Labels

Class Label	Description	Percentage share
BACKGROUND	The citing paper in consideration provides relevant information or is part of the body of literature in this domain	54.93
COMPARES_CONTRASTS	The citing paper expresses similarities or differences to, or disagrees with, the paper is cited	12.27
EXTENSION	The citing paper extends the data, methods etc. of the cited paper.	5.7
FUTURE	The citing paper is potential avenue for future work.	2.07
MOTIVATION	The citing paper is directly motivated by the cited paper.	9.2
USES	The citing paper uses the methodology or tools created by the cited paper.	15.83

Table 2: Purpose Class Labels

Hyperparameter	Parameter value
Embedding dimension	500
# Hidden layer neurons	128
Activation at Bi-LSTM	ReLU
Dropout	0.1
Classifier	Softmax
Loss	Crossentropy
Optimizer	Adam
Learning rate	0.01

Table 3: Hyperparameters used for implementing Bi-LSTM model

Model	Private	Public
Bi-LSTM + D + CW + 128	0.21358	0.18369
Bi-LSTM + D + CW + 64	0.18225	0.17731
Bi-LSTM + D	0.13531	0.16323
RF + CW + TFIDF-SVD	0.12945	0.14014

Table 4: Private and public scores of the submitted models for subtask-A

Model	Private	Public
Bi-LSTM + D + CW + 128	0.48153	0.47758
Bi-LSTM + D + CW + 64	0.47516	0.54010
Bi-LSTM + D	0.46144	0.48346
RF + CW + Fasttext	0.53398	0.47169
RF + TFIDF-SVD	0.43160	0.46971

Table 5: Private and public scores of the submitted models for subtask-B

subtask-A and subtask-B and used the same models for the classification. We experimented with different Bi-LSTM model by addressing the class imbalance problem, which is a problem to be considered, especially in subtask-A. A Bi-LSTM model with dropout and 128 hidden layer neurons achieved the best performance in both tasks. However, random forest classifier with fasttext embedding obtained the highest public score in subtask-B.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Vineetha Rebecca Chacko, M Anand Kumar, and KP Soman. 2019. Experimental study of gender and

language variety identification in social media. In *Advances in Big Data and Cloud Computing*, pages 489–498. Springer.

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

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Suchetha N Kunnath, David Pride, Bikash Gyawali, and Petr Knoth. 2020. Overview of the 2020 wosp 3c citation context classification task. In *Proceedings of the 8th International Workshop on Mining Scientific Publications*, pages 75–83. Association for Computational Linguistics.

Suchetha N Kunnath, David Pride, Drahomira Herrmannova, and Petr Knoth. 2021. Overview of the 2021 sdp 3c citation context classification shared task. In *Proceedings of the Second Workshop on Scholarly Document Processing*.

Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomforest. *R news*, 2(3):18–22.

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

B Premjith, KP Soman, M Anand Kumar, and D Jyothi Ratnam. 2019a. Embedding linguistic features in word embedding for preposition sense disambiguation in english—malayalam machine translation context. In *Recent Advances in Computational Intelligence*, pages 341–370. Springer.

Bhavukam Premjith, Kutti Padannayl Soman, and Prabaharan Poornachandran. 2019b. Amrita_cen@ fact: Factuality identification in spanish text. In *IberLEF@ SEPLN*, pages 111–118.

R Reshma, V Sowmya, and KP Soman. 2018. Effect of legendre–fenchel denoising and svd-based dimensionality reduction algorithm on hyperspectral image classification. *Neural Computing and Applications*, 29(8):301–310.

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

Lina Zhou, Uchechukwuka Amadi, and Dongsong Zhang. 2020. Is self-citation biased? an investigation via the lens of citation polarity, density, and location. *Information Systems Frontiers*, 22(1):77–90.