Bangla News Classification using Graph Convolutional Networks

Conference Paper · April 2021

DOI: 10.1109/ICCCI50826.2021.9402567

3 authors:

Md. Mahbubur Rahman
Iowa State University
20 PUBLICATIONS **142** CITATIONS
SEE PROFILE

Md Akib Zabed Khan
Florida International University
5 PUBLICATIONS **17** CITATIONS
SEE PROFILE

Al Amin Biswas
Bangabandhu Sheikh Mujibur Rahman University, Kishoreganj
36 PUBLICATIONS **319** CITATIONS
SEE PROFILE

2021 International Conference on Computer Communication and Informatics (ICCCI -2021), Jan. 27 – 29, 2021, Coimbatore, INDIA

Bangla News Classification using Graph Convolutional Networks

Md. Mahbubur Rahman
Software Engineer
Crowd Realty
Tokyo, Japan
mahbuburrahman2111@gmail.com

Md. Akib Zabed Khan
Department of CSE
Bangladesh University of Business and Technology
Dhaka, Bangladesh
akibcseju21@gmail.com

Al Amin Biswas
Department of CSE
Daffodil International University
Dhaka, Bangladesh
alaminbiswas.cse@gmail.com

2021 International Conference on Computer Communication and Informatics (ICCCI) | 978-1-7281-5875-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCI50826.2021.9402567

***Abstract*—Online Bangla news has rapidly increased in the era of the information age. Each news site has its different categorization for grouping their news. The Layout and categorization of the online Bangla news articles cannot perpetually meet the individual's needs due to the heterogeneity. So, overcoming this issue and classifying the online Bangla news articles according to the preference of the user is an arduous task. So, it is essential to provide state-of-the-art solutions as well as the best way to solve this problem. The paper aims to build an automated system to classify the Bangla news contents and also find out the state-of-the-art solutions for the small size dataset. It is known that most of the machine learning models need huge amounts of data for the proper training and testing of the models. But due to the scarcity of the dataset, it is not always possible to provide the state-of-the-art solutions. But, in this research work, we have found that Text-GCN performed better than the BiLSTM, GRU-LSTM, LSTM, Char-CNN, and BERT to classify the online Bangla news in spite of the small size of the dataset. The obtained experimental result shows the efficiency of the Text-GCN over the other models in terms of accuracy, precision recall, and F1-score.**

***Keywords*—** Bangla News, Graph Convolutional Networks, Document Classification, NLP.

I. Introduction

Bangladesh is one of the largest populated countries in the world. A major portion of people in this country is very much interested to go through the Bangla news from the internet. In Bangladesh [1], there are thousands of online Bangla news sites continuously presenting the most updated news and information about various issues. Each of the online Bangla news sites has its style, layout, and categorization techniques for presenting their news contents. Due to the diverse issue, this is not always possible to meet the readers' needs. As almost all the online Bangla newspapers are publicly available, it is easy to deal with this problem by further analysis of the data. To extract the huge volume of data from the public website, web scraping techniques can be employed. Web scraping or web harvesting is a technique that is used for extracting data from websites [2].

Some of the work on Bangla documents has already been done by employing various techniques. To classify the Bangla news, Naive Bayes is applied [3]. Based on Inverse Class Frequency [4], the classification of Bangla text documents is performed. To categorize Bengali news [5], several baseline machine learning techniques (Naive Bayes, Random Forest, Logistic Regression, and Linear SVM), and deep learning (Bi-LSTM and CNN) are applied. Decision Tree (C4.5), Na¨ıve Bays (NB), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) are applied to categorize the Bangla web documents [6]. Three popular supervised techniques [7] namely SupportVector Machine (SVM), Na¨ıve Bayes (NB), and Stochastic Gradient Descent (SGD) are applied to categorize Bengali documents.

To get the best performance of the machine learning models, a huge amount of data for the proper training and testing of the model. Because of the scarcity of the dataset, it is essential to investigate and analyze the working capability of different machine learning techniques on small size dataset which is another main purpose of this research work besides classifying the Bangla news.In this research work, we have used BiLSTM, GRU-LSTM, LSTM, Char-CNN, BERT, and Text-GCN to classify the online Bangla news. To evaluate the performance of the working models, we have calculated the accuracy and also other performance evaluation metrics: precision, recall, and F1-score. It is found that Text-GCN outperformed the other models to classify the online Bangla news in spite of the small size of the dataset in terms of accuracy, precision, recall, and F1-score.

For the Text-GCN, we have created a graph network where nodes are representing words and documents. Here, the Adjacency matrix is used to maintain higher neighborhood information among words and documents. Then the graph network is passed to two layers of GCN to classify different Bangla news. We have used a dataset which has six classes namely Education, Economy, International, Entertainment, Sports, and Technology. The employing dataset is collected by web scraping from one of the most popular Bangla news portals namely "Prothom Alo".

This paper is arranged as follows: Section II presents the related study to find the existing research gap. The methodology to accomplish this work is presented in Section III. The experimental result with the details discussion is shown in Section IV. Lastly, Section V concludes the paper by mentioning future works.

II. Literature Review

In this segment, we are portraying some of the related works that have already classified different text data, different news of different languages using different models. There are plenty of works where text classification focuses on linguistic patterns, sentence structure, semantic meaning of the sentences, and some keywords are extracted to detect any specific class. Recurrent Neural Networks and Convolutional Neural Networks are the most used approaches for doing NLP tasks like text classification. A Very Deep Convolutional Neural Networks (VD-CNN) is introduced for text classification

978-1-7281-5875-4/21/$31.00 ©2021 IEEE

single pooling layer for processing several text data. They worked with different text datasets like AG's news, Sogou news, DBPedia, Amazon review full, etc. to categorize English news, Chinese news, and ontology classification, and for some sentiment analysis tasks. A textual-similarity based approach to classify news-related tweets was presented in [9]. O. Demirsoz et al. proposed a model using the binary classification approach: Naive Bayes, libSVM, ADTree algorithms, anomaly detection, and textual similarity methods and classified Turkish tweets collected from Twitter. Their model classified a wide range of news-related topics, but didn't give specific categorical names of these topics. There are some works already done to classify documents in the Bengali language. A character level deep learning approach was presented in [14]. M. M. Rahman et al. used Convolutional Neural Networks (CNN) and Long Short Term Memory (LSTM) to classify Bangla documents encoding the documents at the character level. They used these algorithms to classify documents of three different datasets. Another document classification model is depicted by using transformer based deep learning models in [15]. BERT and ELECTRA models were used there to classify Bangla documents.

A. N. Chy et al. [3] proposed a model to classify Bangla news using the Naive Bayes classifier to categorize 34 concepts of news. They used their own web crawler to collect a dataset from an online news portal and applied the Naive Bayes statistical approach for classifying different news. But their Recall-Precision graph showed an accuracy of less than 80% and didn't give the names of news categories properly. M. R. Hossain et al. [5] presented a model with good accuracy to classify Bangla news using different Machine Learning algorithms and deep learning models. They got higher accuracy using CNN and Bi-LSTM to categorize 12 types of news articles where the dataset is collected from an open-source Bengali corpus of a thesis project. They showed a comparative analysis of getting different accuracy applying 4 different Machine Learning algorithms and Deep learning models: Bi-LSTM and CNN to classify news articles. Machine Learning algorithms and Deep learning approaches are also presented in [10] to classify Chinese news articles. C. M. Huang et al. used Word2vec-SVM, NB, TFIDF-SVM, LSTM, CNN, and Bi-LSTM models and classified different Chinese news of five broad categories: finance, life, politics, sports, and society.

Recently, some text classification tasks and sentiment analysis are done by using Graph Convolutional Networks (GCN). D. Ghosal et al. in [11] built a model to recognize emotion in the conversation using Graph Convolutional Networks. Their proposed DialogueGCN model is capable of recognizing emotion in conversation based on context modeling where two types of contexts are speaker-level context and sequential context. Their dataset contained visual, acoustic, and textual information for all utterances of each and every conversation. The best accuracy to classify emotions for understanding contexts was found by using DialogueGCN among different algorithms like CNN, Memnet, bc-LSTM, DialogueGCN, etc. A web service classification model was presented based on GCN in [12]. H. Ye et al. worked to detect network structure among words of textual web documents and classify them using heterogeneous graph networks. They showed superior accuracy to classify 50 categories of web services implementing the GCN model rather than using TF-IDF+LR, LSTM, LDA Topic Modeling, Bi-LSTM, etc.

From the above studies, we have learned that there are many Machine Learning algorithms and Deep Learning models to classify different text data, and Graph convolutional Networks can give superior accuracy to classify text data. So, we have proposed a model to classify Bangla news into six broad categories implementing Text-GCN and identify the context structure of different types of news. We have got higher accuracy using GCN rather than using other deep learning models to classify Bangla news documents.

III. METHODOLOGY

The methodology section is partitioned into two subsections namely Data Collection and Preprocessing, and News Classification with GCN. The details of the subsections are presented below.

A. Data Collection and Preprocessing

For classifying different news articles, our first task has been to collect a dataset and get different categorical news with particular class names. We have collected data by web scraping from a popular online Bangla news portal named "Prothom Alo". We collected 2400 data for each of the classes. After collecting our dataset, we have performed preprocessing on the collected dataset by following some steps. Firstly, we have removed Bangla stop words that are unnecessary to recognize a category. Then, we have removed unnecessary punctuations and discarded some common words which are not important to classify news categories. We only took the words whose frequency is more than two over the whole data corpus. This helped us to get a dense graph. After cleaning our datasets, we have shuffled our dataset so that all types of articles can be uniformly distributed in all over the dataset. Then we have partitioned our dataset into the train, test, and validation set. 20% (480 data for each individual class) of the dataset is used for the testing purposes. Rest of the 80% dataset is used for the training and validation of the models. Among the 80% (1920 data for the individual class), only 10% (192 data for the individual class) of the dataset is used for the validation of the models.

B. News Classification with GCN

In this task, we have trained and tested our data using graph convolutional networks, a.k.a. GCN. A GCN is one kind of neural network that works on graphs. Given a graph G= (V, E) of N nodes where V and E are the node and edge set respectively. A GCN takes an adjacency matrix A of G and a feature matrix ($X = N_{\times} F^{0}$) of the nodes as input and produces output features ($O = N_{\times} F^{L}$) for each node. The diagonal elements of A are set to 1 to add the node's own feature during propagation. GCN can propagate features to the one-hop neighbors with one-layer convolution. Larger neighborhood information can be propagated through stacking multiple GCN layers. The propagation rule for a GCN network is as follows equation (1) and (2) [13]:

$$\tilde{A} = D^{(-1/2)} A D^{(-1/2)} \tag{1}$$

$$L^{(j+1)} = \sigma(\tilde{A} L^{(j)} W_j) \tag{2}$$

Where D, $\tilde{A}$, W_j, j, L are the degree matrix, normalized symmetric adjacency, weight matrix, layer number and output of a GCN layer respectively. $L^{(j)}$ is same as the input feature(X) when j = 0 and same as the output of the previous layer when j > 0.

To build the graph, we have considered both words and documents as nodes. So, the number of nodes in our graph is the summation of the number of documents and the number of unique words in the dataset. Also, the initial input feature X is set to I which is the one-hot vector of the nodes. We set the weight of an edge between a word and a document by the TF-IDF of the word in the document. On the other hand, the edge between two words is calculated based on their co-occurrences. A fixed-size sliding window is used to achieve the co-occurrence statistics over the dataset. To calculate the weight between two nodes, we have applied pointwise mutual information[13]. Formally, the adjacency matrix is built by setting weight by the following formulas (eq. 3 to 6):

$$A_{ij} = \begin{cases} TF - IDF_{ij}, & \text{where i is document, j is word} \\ PMI(i, j), & \text{where i, j are words, PMI(i, j) > 0} \\ 1, & \text{where i = j} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

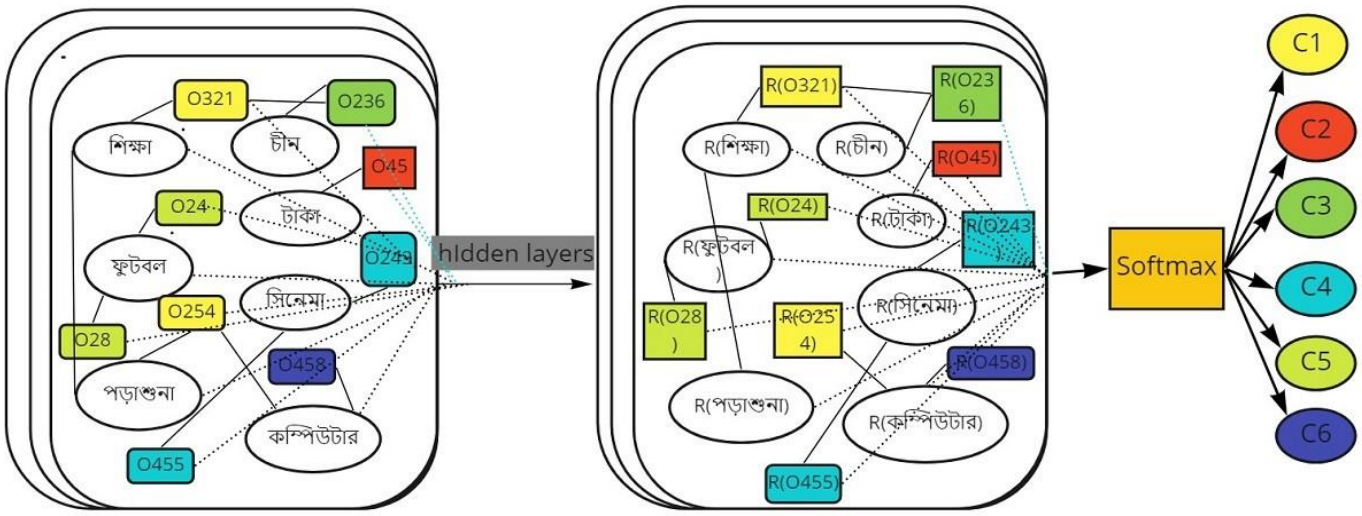

Fig. 1. Schema diagram of Text-GCN for our dataset. Document nodes are represented by starting with an "O" letter, others nodes mean word nodes. R(x) symbolizes the embedding of x. Various document classes are represented by different colors(six colors are shown) and C1 to C6 represent the output classes.

where

$$PMI(i,j) = \frac{log(p(i,j))}{p(i) * p(j)} \quad (4)$$

$$p(i,j) = \frac{SW(i,j)}{SW} \quad (5)$$

$$p(i) = \frac{SW(i)}{SW} \quad (6)$$

Here, SW(i) is the quantity of sliding windows possessing word i, SW(i, j) is the quantity of sliding windows having both words i and j, and SW is the total sliding windows in the dataset.

If the PMI value is negative, then we did not add any edge between these words as a negative value implies a very low semantic correlation between these words in the corpus. No edge is added between two documents. Finally, we fed the text graph and the input feature (X) to a two-layer GCN (eq. 7 to 8). The final layer's output features(O) that have equal size as the labels are sent to a softmax layer (eq. 9) to classify the documents.

$$L^1 = ReLU(\tilde{A}XW_0) \quad (7)$$

$$O = L^2 = \tilde{A}L^1W_1 \quad (8)$$

$$Y = softmax(O) \quad (9)$$

Where $\tilde{A}$ is same for each layer, and Y is the final output value based on TABLE I.

TABLE I
MULTI-LABEL CLASS FOR THE DATASET

Encoded Value (Y)	Output Class
0	Education (C1)
1	Economy (C2)
2	International (C3)
3	Entertainment (C4)
4	Sports (C5)
5	Technology (C6)

The overall schematic illustration of text GCN is shown in Fig. 1.

To train the model, we have used cross-entropy function as a loss function. The weight W_0 and W_1 are trained by gradient descent. Overall, this two-layer GCN can propagate features within two hop distances. We have trained and tested by adding more layers but it does not improve the accuracy.

IV. RESULT AND DISCUSSION

To evaluate the applied models, we have calculated accuracy and performance evaluation metrics, namely precision, recall, and F1-Score. These three metrics are very well known to measure the performance of the models. The following formulas (Eq. 10 to 13) are used to calculate the value of these metrics.

$$Accuracy: A = \frac{TP + TN}{TP + TN + FN + FP} * 100\% \quad (10)$$

$$Precision: P = \frac{TP}{TP + FP} * 100\% \quad (11)$$

$$RecallRate: R = \frac{TP}{TP + FN} * 100\% \quad (12)$$

$$F1\_Score: F1 = \frac{2 * P * R}{P + R} * 100\% \quad (13)$$

Accuracy increases over the epochs and loss decreases over the epochs are the fundamental behavior of a good model. Training and validation accuracy over the epochs are presented by Fig. 2, and the model's training and validation loss over the epochs are represented by Fig. 3. These figures demonstrate that the accuracy of the Text-GCN increases over the epochs where loss decreases over the epochs, which is a good sign for a technique.

To evaluate the quality of the model's output, the Receiver Operating Characteristic (ROC) curve is usually used. A larger area under the ROC Curve (AUC) usually indicates the better performance of the model. The ROC curve of Text-GCN is represented in Fig. 4. From Fig. 4, it is observed that the ROC of Text-GCN has enough larger AUC (Area is 0.98 for both micro and macro average ROC Curve) which indicates that the applied model is working better here. Micro-averaging and macro-averaging are the two measures for multi-label classification. Here, we have considered both micro and macro averaging techniques for this work.

TABLE II shows the class-wise evaluation metrics results of Text-GCN. Here, the metrics precision, recall, and F1-Score are presented.

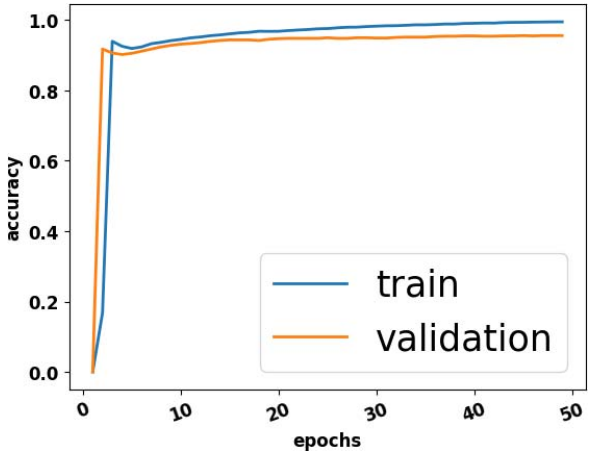

Fig. 2. Training and Validation Accuracy loss over the epochs of Text-GCN

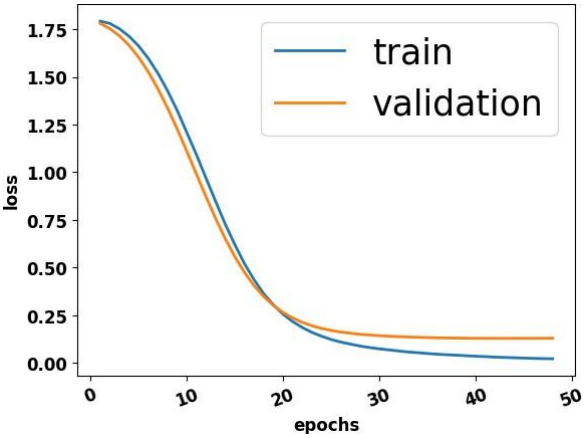

Fig. 3. Training and Validation Accuracy loss over the epochs of Text-GCN

For the class "Sports", the precision, recall, and F1-score of Text-GCN are 99.38%, 99.58%, and 99.48%, respectively, which is the highest among all presented.

TABLE II
CLASS-WISE EVALUATION METRICS RESULTS OF TEXT-GCN.

Category	Precision	Recall	F1-Score
Economy	97.48%	96.88%	97.18%
Education	96.07%	96.88%	96.47%
Entertainment	96.25%	96.26%	96.25%
International	94.62%	95.21%	94.91%
Sports	99.38%	99.58%	99.48%
Technology	93.68%	92.71%	93.19%

In TABLE III, we have presented text GCN's performance with some other state of the arts models. From the presented data, it is found that Text-GCN outperformed all the other applied models in terms of precision, recall, F1-score, and accuracy. Text-GCN achieved 96.25% precision, 96.25% recall, 96.25% F1-Score, and 96.25% accuracy.

V. Conclusion

The aim of this research work is to classify the online Bangla news and find the state-of-the-art solutions for classifying the Bangla news

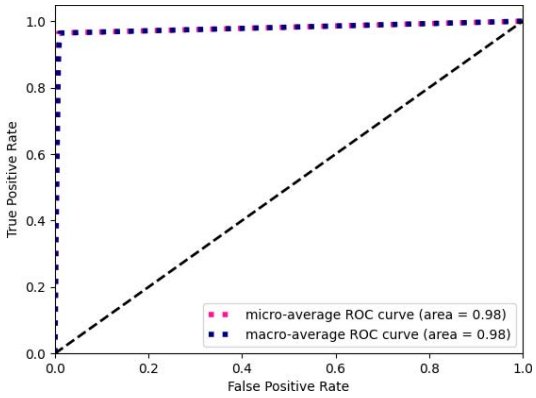

Fig. 4. Roc Curve of Text-GCN.

TABLE III
COMPARISON OF TEXT-GCN WITH DIFFERENT STATE-OF-THE-ART MODELS .

Model	Precision	Recall	F1-Score	Accuracy
BiLSTM	92.84%	92.81%	92.80%	92.81%
GRU-LSTM	95.04%	95.00%	95.01%	95.00%
LSTM	94.44%	94.44%	94.44%	94.44%
Char-CNN	86.86%	86.84%	86.82%	86.84%
BERT	94.96%	94.97%	94.95%	94.97%
Text-GCN	96.25%	96.25%	96.25%	96.25%

with the small dataset. The dataset which is collected by web scraping from the official website of Prothom Alo is considered to accomplish this work. Here, 80% of data is used for the model's training and validation. The remaining 20% data is used for testing purposes. In this paperwork, we have mainly worked with Text-GCN to classify the Bangla documents. To compare the performance of Text-GCN with other state-of-the-art techniques to classify the online Bangla news where we have applied BiLSTM, GRU-LSTM, LSTM, Char-CNN, and BERT besides the Text-GCN. From the result comparison, it is found that Text-GCN achieved the highest accuracy which is 96.25%. The precision, recall, and F1-score of Text-GCN are highest also and which are 96.25% in each section. As the Text-GCN model consumes a large amount of memory, we have to use a small amount of data to train our model. As a larger dataset leads to higher accuracy in deep learning-based models, we will try to solve the memory related problems in our future work. So, we will use other machine learning and deep learning techniques while considering more classes in the dataset.

References

[1] "All Bangla News Paper," Available Online: http://www.bdallbanglanewspaper.com/ [Last Accessed 02 October 2020]

[2] "Web Scraping," Available Online: https://en.wikipedia.org/wiki/Web_scraping [Last Accessed 02 October 2020]

[3] A. N. Chy, M. H. Seddiqui and S. Das, "Bangla news classification using naive Bayes classifier," 16th Int'l Conf. Computer and Information Technology, Khulna, 2014, pp. 366-371, doi: 10.1109/ICCITechn.2014.6997369.

[4] A. Dhar, N. S. Dash and K. Roy, "Classification of Bangla Text Documents based on Inverse Class Frequency," 2018 3rd International Conference On Internet of Things: Smart Innovation and Usages (IoT-SIU), Bhimtal, 2018, pp. 1-6, doi: 10.1109/IoT-SIU.2018.8519866.

[5] M. R. Hossain, S. Sarkar, and M. Rahman, "Different Machine Learning based Approaches of Baseline and Deep Learning Models for Bengali News Categorization," International Journal of Computer Applications, 975, 8887.
[6] A. K. Mandal and R. Sen, "Supervised learning methods for Bangla web document categorization", International Journal of Artificial Intelligence Applications (IJAIA), Vol. 5, No. 5, 2014.
[7] M. Islam, F. E. M. Jubayer, and S. I. Ahmed, "A comparative study on different types of approaches to Bengali document categorization," arXiv preprint arXiv:1701.08694 (2017).
[8] A. Conneau, H. Schwenk, L. Barrault, Y. Lecun, "Very deep convolutional networks for text classification.", arXiv preprint arXiv:1606.01781., Jun 6, 2016.
[9] O. Demirsoz, R. Ozcan, "Classification of news-related tweets.", Journal of Information Science, 43(4):509-24, Aug 2017.
[10] C. M. Huang, Y. J. Jiang, "An Empirical Study on the Classification of Chinese News Articles by Machine Learning and Deep Learning Techniques." In: 2019 International Conference on Machine Learning and Cybernetics (ICMLC), IEEE, (pp. 1-6), Jul 7, 2019.
[11] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation.", arXiv preprint arXiv:1908.11540, Aug 30, 2019.
[12] H. Ye, B. Cao, J. Chen, J. Liu, Y. Wen, J. Chen, "A Web Services Classification Method Based on GCN." In: 2019 IEEE Intl Conf on Parallel Distributed Processing with Applications, Big Data Cloud Computing, Sustainable Computing Communications, Social Computing Networking (ISPA/BDCloud/SocialCom/SustainCom), IEEE, (pp. 1107-1114), Dec 16, 2019.
[13] L. Yao, C. Mao, Y. Luo, "Graph convolutional networks for text classification.", In: Proceedings of the AAAI Conference on Artificial Intelligence, (Vol. 33, pp. 7370-7377), Jul 17, 2019.
[14] MM Rahman, R. Sadik and A. A. Biswas, "Bangla Document Classification using Character Level Deep Learning,", 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Istanbul, Turkey, pp. 1-6, 2020.
[15] MM Rahman, MA Pramanik, R. Sadik, M. Roy and P. Chakraborty, "Bangla Documents Classification using Transformer Based Deep Learning Models. In: Proceedings of the 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 19-20 December, 2020.