2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON) | 978-1-6654-0690-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/UEMCON53757.2021.9666633

Review of Graph Neural Network in Text Classification

Masoud Malekzadeh
Department of Electrical and Computer Engineering
University of Massachusetts at Lowell
Lowell, USA
masoud_malekzadeh@student.uml.edu

Parisa Hajibabaee
Department of Computer Science
University of Massachusetts at Lowell
Lowell, USA
parisa_hajibabaee@student.uml.edu

Maryam Heidari
George Mason University
Virginia, USA
mheidari@gmu.edu

Samira Zad
Knight Foundation School of CIS
Florida International University
Miami, FL
szad001@fiu.edu

Ozlem Uzuner
George Mason University
ouzuner@gmu.edu

James H Jr Jones
George Mason University
jjonesu@gmu.edu

***Abstract*—Text classification is one of the fundamental problems in Natural Language Processing (NLP). Several research studies have used deep learning approaches such as Convolution Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for text classification. Over the past decade, graph-based approaches have been used to solve various NLP tasks including text classification. This paper reviews the most recent state-of-the-art graph-based text classification, datasets, and performance evaluations versus baseline models.**

***Index Terms*—graph neural network, graph convolutional network, attention layer, deep learning, text Classification**

I. Introduction

Text classification is one of the fundamental problems in Natural Language Processing (NLP), and it has a wide range of applications in a variety of fields, including social media, healthcare, and bioinformatics. Several approaches have been developed to address this problem. Classical text classification methods mainly focus on hand-crafted feature engineering and classification methods. For example, [1] proposed a visualization approach that reduces cognitive load and increases the speed of text labeling. One of the most straightforward approaches to represent natural language is bag-of-words model. However, such an interpretation of language is oblivious to the order and placement of the tokens in the text and only examines the frequency of the tokens in the text. A more advanced model is to consider the relationship among the words in a text besides their frequencies. The model takes into account the order of tokens in a sentence and text. One advantage of this representing language in an ordered sequence of words is its ability to capture more textual information. Due to its great power in modeling non-Euclidean, graphs have become ubiquitous in the NLP domain. Compared to the other categories, graphs can capture more complex interactions among the tokens in the text.

This paper is organized as follows: A back-

978-1-6654-0690-1/21/$31.00 ©2021 IEEE

ground on graph neural networks is presented in section II. An overview of the current state of the are works are presented in section III. Finally, the conclusion is given in Section IV.

II. Graph Neural Network

Graphs are abstract data structures that can be defined using an adjacency matrix, indicating whether two nodes are connected. If the graph is unweighted, the value associated with that node in the adjacency matrix will be 0 if there is no connection and 1 otherwise. To be formal, let us consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents a set of nodes and $\mathcal{E}$ represents a set of edges between these nodes. An adjacency matrix $\mathbf{A} \in \mathbb{R}^{|V| \times |V|}$ can represent the existing edge between nodes. In graph every node is connected to itself. $X \in \mathbb{R}^{n \times m}$ is a matrix containing the nodes with their feature vectors with length m. Graph Neural Networks (GNNs) are a type of deep learning method that can be applied directly to graphs to make inferences on data provided in a graph. They make it simple to perform node-level, edge-level, and graph-level prediction tasks. Deep learning has gotten a lot of attention in the last decade, especially in problem domains including classification, detection, and verification [2]–[4]. It is worth noting that popular deep learning techniques like *Convolutional Neural Networks (CNNs)*, widely used in healthcare area [5]–[9] and *Recurrent Neural Networks (RNNs)* do not apply for graph-structured data. Although Deep Learning algorithms have shown remarkable progress in many domains such as medical image enhancement [10]–[12], we need to create a new type of deep learning architecture to construct a deep neural network across general graphs. Graph Convolutional Network (GCN) [13] is a simple and practical variant of GNN that can capture high-order neighbor node's information. The goal of GCN is to implement mapping nodes to a $d-$dimensional embedding space so that similar nodes in the graph are embedded close to each other. However, in recent years, architectures that use the graph aspect of text as an input have outperformed traditional word embeddings and automatic feature extraction. Machine learning models have different applications in text classification in different scientific fields [14]–[19] [20]–[27].

III. Survey

Table I summarized the recent state-of-the-art graph-based text classification methods along with the baseline models for comparison and datasets.

A. Text GCN

Yao et al. [32] used a GCN for text classification. To model global word co-occurrence, they employed a heterogeneous text graph with word nodes and document nodes. In their model, the number of nodes is determined by the number of documents and unique words. Edges between nodes are built using word occurrence in documents and word co-occurrence across the corpus. The frequency-inverse document frequency determines the weight of the edge between a document and a word node (TF-IDF). The prepared graph is fed into a two-layer GCN that allows message passing among the nodes two-step away at maximum. Based on test results, a two-layer GCN has a better performance than one layer, and increasing the number beyond two layers does not increase the model's accuracy.

For evaluation of the performance of the proposed model, the Text GCN was compared with the following state-of-the-art architectures: Bag-of-words model with term frequency-inverse document frequency weighting (TF-IDF+LR) [33], Convolutional Neural Network (CNN) [28], Long Short Term Memory (LSTM) [34], Bi-directional LSTM (Bi-LSTM) [35], Paragraph Vector Model(PV-DBOW - PV-DM) [36], predictive text embedding (PTE) [37], FastText [38], Simple Word Embedding Models (SWEM) [39], Label Embedding Attentive Models

TABLE I
RECENT STATE-OF-THE-ART GRAPH-BASED TEXT CLASSIFICATION METHODS ALONG WITH THE BASELINE MODELS FOR COMPARISON AND DATASETS.

Name	Architecture	Baseline Models	Data Set	Results
Yao et al. [28]	Two layer GCN	TF-IDF + LR LSTM Bi-LSTM PV-DBOW PV-DM PTE fastText SWEM LEAM Graph-CNN-C Graph-CNN-S Graph-CNN-F	20NG Ohsumed R52 R8 MR	Text GCN outperforms all baseline models in all datasets
Huang et al. [29]	Single GCN+Softmax	CNN LSTM fastText Graph-CNN Text-GCN	R8 R52 Ohsumed	GNN model outperforms traditional methods (CNN LSTM fastText graph-based approaches)
Pal et al. [30]	Single Graph attention layer + BiLSTM + BERT	HSVM RCNN HAN Bi-BloSAN DCNN SGM+GE HR-DGCNN TEXTCNN HE-AGCRCNN HTrans BOW-CNN HilAP BERT	Reuters-21578 RCV1-V2 Arxiv Academic Paper Slashdot Toxic comment	20% improvement compared to HSVM 11%, 19%, 5%, and 8% accuracy improvement compared to HE-AGCRCNN, HAN, HiLAP, HTrans 16% improvement compared to Bi-BloSAN 12% and 6% accuracy compared to TEXTCNN and BOW-CNN 2% improvement over BERT model
Tayal et el [31]	Two layer GCN + Softmax	TF-IDF + LR CNN-rand CNN-non-static CharCNN fastText Graph-CNN-C Text GCN SWEM	Amazon Internal Electronics Home and kitchen	5.86% increase in accuracy over SWEM (Amazon dataset) 1.8% and 2.4% performance gain over (electronics and home kitchen datasets)

(LEAM) [40], Graph CNN Model (Graph-CNN-C) [41], Graph CNN using spline filter (Graph-CNN-S) [42], Graph CNN using fourier filter (Graph-CNN-F) [43]. For evaluating the performance, the Text GCN was trained and validated using five data sets: 1. 20-Newsgroups (20NG) 2. Ohsumed 3. R52 4.R8 of Reuters 5. Movie Review (MR) data set.

Based on experimental results, the Text GCN outperformed the other baseline models. There are two main reasons for the performance of the Text GCN: 1. Text GCN is capable of capturing document-word and global word-word relationships. 2. The Text GCN calculates the new features as a weighted sum of itself and its second-order neighbors. However, the Text GCN fails to outperform CNN and LSTM models on the MR data set because GCN does not consider word orders which is crucial in sentiment classification.

B. *Text Level GNN for Text Classification*

Huang et al. [29] proposed a novel GNN approach for text classification that generates graphs for each input text while sharing global parameters. This technique drops the dependency between an individual text and the whole corpus, allowing for online testing while preserving global information. Moreover, graphs with smaller text windows can capture more local information and dramatically decrease edge counts and memory usage.

To build the word graph, all words in a text are regarded as nodes of the graph, and they share an edge with the words in the text that are next to them. As a result, the created graph has fewer edges and nodes than GNN models, which reduces GPU memory consumption. The representations of the nodes are updated using their own information as well as information acquired from neighboring nodes as context representation utilizing the Message passing approach. The model was tested and compared against a number of other models, including CNN [28], LSTM [34], fastText [38], Graph-CNN [41], Text-GCN [32] on the R8, R521, and Ohsumed2 data sets.

According to experimental data, the GNN model outperforms conventional text classification methods because the graph structure allows nodes to learn a more accurate representation compared to the traditional ones. It has a better performance compared to other graph-based text classification models like GraphCNN. The main reason for such a higher performance is because GraphCNN cannot distinguish the importance of words since it represents the document as a bag of words using a big window with no weighted edges. In contrast to GraphCNN, the model used in this work employs trainable edge weights that are shared globally and learned using the whole corpus text, allowing words to express themselves differently when faced with different collocations. According to experimental data, the GNN employed in this study consumes less memory than the Text-GCN and has fewer edges. One of the key reasons for this model's benefit in terms of memory consumption is that there are no text nodes in this model since they are calculated as the sum of the representations of the word nodes, which reduces memory consumption. Moreover, in contrast to Text-GCN, which employs a larger window to collect co-occurrence in the whole corpus to obtain a more accurate weight, the words in this model are only connected to adjacent nodes. As a result, the edge weight matrix will become more sparse than it was previously.

C. MAGNET: Multi-Label Text Classification Using Attention-based GNN

Multiclass Label Text Classification (MLTC) is the task of assigning different labels to a document in the corpus and has many applications such as text categorization, tag recommendation, and information retrieval [44]. Several approaches for tackling the MLTC problems have been presented in the literature [45] [46]. However, existing classical, deep learning, and GNN-based MLTC approaches [47] [48] fail to capture the correlation among the labels. To better capture the connection between the labels, Pal et al. [30] introduced a unique architecture called Multi-label Text Classification using Attention-based Graph Neural Network (MAGNET) to overcome the above problem by automatically learning the relationships between labels based on the feature matrix.

The MAGNET aims to learn the adjacency matrix of a graph in order to model the correlation of the labels as the node of the graph. The key premise in this MAGNET is that by modeling the correlation among labels as a weighted graph, the Graph Attention network (GAT) can learn the adjacency matrix and the attention weights simultaneously to represent the correlation among the labels.

In MAGNET, the embedding vectors of the labels perform as the node features, and the adjacency matrix is a learnable parameter. The attention mechanism in MAGNET can determine the importance of labels in a correlation graph by learning the importance of their adjacent labels. MAGNET uses a cascade of multi-head attention layers [49] to describe labels relationships. The input to the first layer is the embedding matrix, and the output from the previous layers is fed into the following layers. MAGNET used the BERT model to obtain the embedding of the words and fed the embeddings to a bidirectional LSTM to get the feature vector. A combination of the three methods was used to initialize the adjacency matrix, including 1. Identity: zero correlation at the start; 2. Xavier Matrix [50] to initialize weights of the matrix; 3. correlation matrix: Counting pairwise co-occurrence of labels. The model was tested and compared against three categories of methods, including 1. Flat baselines 2. Sequence graph and N-gram-based models 3. Recent state-of-the-art models on the R8, R521, and Ohsumed2 data sets.

Based on experimental results, MAGNET significantly outperformed comparing the state-of-the-art architectures, including 20% improvement compared to the HSVM model, 11% , 19% , 5% , and 8% accuracy improvement compared to HE-AGCRCNN, HAN, HiLAP, HTrans respectively. Furthermore, compared with the best Hierarchical text classification models, a 16% improvement in miF1 over the bi-directional block self-attention network (Bi-BloSAN). For CNN models, the MAGNET achieved a 12% and 6% accuracy improvement compared with TEXTCNN and BOW-CNN methods, respectively, and a 2% accuracy improvement over the state-of-the-art BERT model. Moreover, the empirical result indicates that applying the attention mechanism in a correlation graph can determine the importance of a label by considering the significance of its neighbor labels and improving the average miF1 score by 4% compared to the GCN model.

D. Short Text

With the increasing growth of social media, e-commerce, and online communication, short text classification has become got attention in NLP. The short text is commonly seen in tweets, chat messages, search queries, product descriptions, and online reviews. Tayal et al. [31] have used the graph convolution network for short text classifications (STGCN). Short text classification is one of the challenging problems in NLP for several reasons. First, a short text is highly sparse and lacks adequate features to offer enough word co-occurrence. Second, unlike other text resources, most short text corpora do not have a language structure or follow grammar rules.

Traditional classification approaches [51] [52] fail to generalize for the classification task [53], and the state-of-the-art deep learning approaches' performance is limited. Tayal et al. [31] have used the graph convolution network for short text classifications. The proposed architecture is composed of a two-layer GCN followed by a softmax layer. The proposed GCN architecture was developed for short text classification for two tasks, Product Title Classification and Product Query classification. The authors investigated GNC's performance on three datasets and compared it to several baseline including TF-IDF + LR [33] [54], CNN-rand [28] , CharCNN [55], fastText [38], Graph-CNN-C [41] ,Text GCN [32], SWEM [39] models to evaluate its performance. According to experimental results, ST-GCN obtained 5.86% increase in accuracy over the second-best baseline model, SWEM, for the Amazon Internal dataset. On the other hand, ST-GCN demonstrated a 1.8% and 2.4% performance gain over the second-best baseline model in the electronics and home kitchen datasets, respectively. Furthermore, for short text datasets, TF-IDF + LR outperforms deep learning models such as CNN-random. The reason is deep neural network effectiveness depends on capturing structure in the corpus, which is scarce in short text texts.

IV. Conclusion

In this paper we have reviewed graph neural and graph convolution papers for text classification task. Based on their evaluation results, they have outperformed the traditional and deep learning based counterpart on the reported data sets. In the future, we will conduct a thorough review of studies that have used graph neural networks in text classification.

References

[1] S. Enayati, Z. Yang, B. Lu, and S. Vucetic, "A visualization approach for rapid labeling of clinical notes for smoking status extraction," in *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances*, 2021, pp. 24–30.

[2] M. Mostofa, S. Mohamadi, J. Dawson, and N. Nasrabadi, "Deep gan-based cross-spectral cross-resolution iris recognition," *arXiv preprint arXiv:2108.01569*, 2021.

[3] S. Mohamadi and H. Amindavar, "Deep bayesian active learning, a brief survey on recent advances," *arXiv preprint arXiv:2012.08044*, 2020.

[4] S. Mohamadi, D. Adjeroh, B. Behi, and H. Amindavar, "A new framework for spatial modeling and synthesis of genomic sequences," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020, pp. 2221–2226.

[5] R. Habibi, "Detecting surface interactions via a wearable microphone to improve augmented reality text entry," 2021.

[6] S. K. Tabbakh, R. Habibi, and S. Vafadar, "Design and implementation of a framework based on augmented reality for phobia treatment applications," in *2015 International Congress on Technology, Communication and Knowledge (ICTCK)*. IEEE, 2015, pp. 366–370.

[7] R. Shults, E. Levin, R. Habibi, S. Shenoy, O. Honcheruk, T. Hart, and Z. An, "Capability of matterport 3d camera for industrial archaeology sites inventory." *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2019.

[8] E. Levin, R. Shults, R. Habibi, Z. An, and W. Roland, "Geospatial virtual reality for cyberlearning in the field of topographic surveying: Moving towards a cost-effective mobile solution," *ISPRS International Journal of Geo-Information*, vol. 9, no. 7, p. 433, 2020.

[9] E. Levin, W. Roland, R. Habibi, Z. An, and R. Shults, "Rapid visual presentation to support geospatial big data processing," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 43, pp. 463–470, 2020.

[10] J. Liu, M. Malekzadeh, N. Mirian, T. Song, C. Liu, and J. Dutta, "Artificial intelligence-based image enhancement in pet imaging: Noise reduction and resolution enhancement," *PET clinics*, vol. 16, no. 4, pp. 553–576, 2021.

[11] D. Bamgboje, I. Christoulakis, I. Smanis, G. Chavan, R. Shah, M. Malekzadeh, I. Violaris, N. Giannakeas, M. Tsipouras, K. Kalafatakis *et al.*, "Continuous non-invasive glucose monitoring via contact lenses: Current approaches and future perspectives," *Biosensors*, vol. 11, no. 6, p. 189, 2021.

[12] M. Heidari and S. Rafatirad, "Semantic convolutional neural network model for safe business investment by using bert," in *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, 2020, pp. 1–6.

[13] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *CoRR*, vol. abs/1609.02907, 2016.

[14] M. Heidari, J. H. Jones, and O. Uzuner, "Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter," in *2020 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2020, pp. 480–487.

[15] M. Heidari and S. Rafatirad, "Using transfer learning approach to implement convolutional neural network model to recommend airline tickets by using online reviews," in *2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA*. IEEE, 2020, pp. 1–6.

[16] M. Heidari, S. Zad, and S. Rafatirad, "Ensemble of supervised and unsupervised learning models to predict a profitable business decision," in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. IEEE, 2021, pp. 1–6.

[17] M. Heidari and J. H. Jones, "Using bert to extract topic-independent sentiment features for social media bot detection," in *2020 11th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, 2020, pp. 0542–0547.

[18] S. Zad, M. Heidari, J. H. Jones, and O. Uzuner, "A survey on concept-level sentiment analysis techniques of textual data," in *2021 IEEE World AI IoT Congress (AIIoT)*. IEEE, 2021, pp. 0285–0291.

[19] S. Zad, M. Heidari, H. James Jr, and O. Uzuner, "Emotion detection of textual data: An interdisciplinary survey," in *2021 IEEE World AI IoT Congress (AIIoT)*. IEEE, 2021, pp. 0255–0261.

[20] M. Heidari, H. James Jr, and O. Uzuner, "An empirical study of machine learning algorithms for social media bot detection," in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. IEEE, 2021, pp. 1–5.

[21] M. Heidari, S. Zad, B. Berlin, and S. Rafatirad, "Ontology creation model based on attention mechanism for a specific business domain," in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. IEEE, 2021, pp. 1–5.

[22] M. Heidari and S. Rafatirad, "Bidirectional transformer based on online text-based information to implement convolutional neural network model for secure business investment," in *2020 IEEE International Symposium on Technology and Society (ISTAS)*. IEEE, 2020, pp. 322–329.

[23] P. Hajibabaee, M. Malekzadeh, , M. Heidari, S. Zad, O. Uzuner, and J. H. Jones, "An empirical study of the graphsage and word2vec algorithms for graph multiclass classification," in *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2021.

[24] S. Zad, M. Heidari, P. Hajibabaee, and M. Malekzadeh, "A survey of deep learning methods on semantic similarity and sentence modeling," in *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 2021.

[25] M. Heidari, S. Zad, M. Malekzadeh, P. Hajibabaee, , S. HekmatiAthar, O. Uzuner, and J. H. J. Jones, "Bert model for fake news detection based on social bot activities in the covid-19 pandemic," in *2021 12th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*. IEEE, 2021.

[26] S. Zad, J. Jimenez, and M. Finlayson, "Hell hath no fury? correcting bias in the nrc emotion lexicon," in *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, 2021, pp. 102–113. [Online]. Available: https://aclanthology.org/2021.woah-1.11

[27] S. Zad and M. Finlayson, "Systematic evaluation of a framework for unsupervised emotion recognition for narrative text," in *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*. Association for Computational Linguistics, 2020, pp. 26–37. [Online]. Available: https://aclanthology.org/2020.nuse-1.4

[28] Y. Kim, "Convolutional neural networks for sentence classification," *CoRR*, vol. abs/1408.5882, 2014.

[29] L. Huang, D. Ma, S. Li, X. Zhang, and H. Wang, "Text level graph neural network for text classification," *CoRR*, vol. abs/1910.02356, 2019.

[30] A. Pal, M. Selvakumar, and M. Sankarasubbu, "Multi-label text classification using attention-based graph neural network," *CoRR*, vol. abs/2003.11644, 2020.

[31] K. Tayal, "Short text classification using graph convolutional network," 2019.

[32] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," *CoRR*, vol. abs/1809.05679, 2018.

[33] M. Das, S. Kamalanathan, and P. Alphonse, "A comparative study on tf-idf feature weighting method and its analysis using unstructured dataset," in *COLINS*, 2021.

[34] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," *CoRR*, vol. abs/1605.05101, 2016.

[35] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," *CoRR*, vol. abs/1508.01991, 2015.

[36] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," *CoRR*, vol. abs/1405.4053, 2014.

[37] J. Tang, M. Qu, and Q. Mei, "PTE: predictive text embedding through large-scale heterogeneous text networks," *CoRR*, vol. abs/1508.00200, 2015.

[38] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *CoRR*, vol. abs/1607.01759, 2016.

[39] D. Shen, G. Wang, W. Wang, M. R. Min, Q. Su, Y. Zhang, C. Li, R. Henao, and L. Carin, "Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms," *CoRR*, vol. abs/1805.09843, 2018.

[40] G. Wang, C. Li, W. Wang, Y. Zhang, D. Shen, X. Zhang, R. Henao, and L. Carin, "Joint embedding of words and labels for text classification," *CoRR*, vol. abs/1805.04174, 2018.

[41] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," *CoRR*, vol. abs/1606.09375, 2016.

[42] J. Bruna, W. Zaremba, A. Szlam, and Y. Lecun, "Spectral networks and locally connected networks on graphs," in *International Conference on Learning Representations (ICLR2014), CBLS, April 2014*, 2014.

[43] M. Henaff, J. Bruna, and Y. LeCun, "Deep convolutional networks on graph-structured data," *CoRR*, vol. abs/1506.05163, 2015.

[44] M. K. Dalal and M. A. Zaveri, "Automatic text classification: a technical review," *International Journal of Computer Applications*, vol. 28, no. 2, pp. 37–40, 2011.

[45] J. D. Rennie and R. Rifkin, "Improving multiclass text classification with the support vector machine," 2001.

[46] J. D. Rennie, "Improving multi-class text classification with naive bayes," 2001.

[47] F. Gargiulo, S. Silvestri, and M. Ciampi, "Deep convolution neural network for extreme multi-label text classification." in *Healthinf*, 2018, pp. 641–650.

[48] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," in *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, 2017, pp. 115–124.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[50] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural net-

works," in *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'10). Society for Artificial Intelligence and Statistics*, 2010.

[51] P. Hajibabaee, F. Pourkamali-Anaraki, and M. Hariri-Ardebili, "An empirical evaluation of the t-sne algorithm for data visualization in structural engineering," *arXiv preprint arXiv:2109.08795*, 2021.

[52] ——, "Kernel matrix approximation on class-imbalanced data with an application to scientific simulation," *IEEE Access*, pp. 83 579–83 591, 2021.

[53] G. Song, Y. Ye, X. Du, X. Huang, and S. Bie, "Short text classification: A survey," *Journal of multimedia*, vol. 9, no. 5, p. 635, 2014.

[54] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, p. 513–523, Aug. 1988.

[55] X. Zhang, J. J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *CoRR*, vol. abs/1509.01626, 2015.