Comparative Study of Word Embeddings for Classification of Scientific Article on Human Health Risk of Electromagnetic Fields

Sang-Woo Lee
Division of Software
Hallym University
Chuncheon, South Korea
sam_2011@hallym.ac.kr

Jung-Hyok Kwon
Smart Computing Laboratory
Hallym University
Chuncheon, South Korea
jhkwon@hallym.ac.kr

Sol-Bee Lee
Division of Software
Hallym University
Chuncheon, South Korea
thfqla3535@hallym.ac.kr

Nam Kim
School of Information and Communication Engineering
Chungbuk National University
Cheongju, South Korea
namkim@chungbuk.ac.kr

Hyung Do Choi
Radio Technology Research Department
Electronics and Telecommunications Research Institute (ETRI)
Daejeon, South Korea
choihd@etri.re.kr

Eui-Jik Kim*
Division of Software
Hallym University
Chuncheon, South Korea
ejkim32@hallym.ac.kr

2022 IEEE 11th Global Conference on Consumer Electronics (GCCE) | 978-1-6654-9232-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/GCCE56475.2022.10014242

***Abstract*—This paper presents a comparative study on pre-trained word embeddings for the classification of scientific articles on the human health risk of EMF. In this study, the various neural network (NN) models are trained using the three biomedical pre-trained word embeddings (i.e., PubMed-word2vec, BioWordVec, and PubMed-BERT). Then, the performances of the trained NN models are evaluated to compare the pre-trained word embeddings. The evaluation results show that the NN models using PubMed-BERT outperform the other NN models using PubMed-word2vec and BioWordVec in the classification of scientific articles on the human health risk of EMF.**

Keywords—Word Embeddings, Deep Neural Network, Human Health Risk of Electromagnetic Fields, Exposure Assessment

I. Introduction

Recently, as the number of consumer electronics using radio frequency (RF) has sharply increased, the adverse effects of exposure to electromagnetic fields (EMF) on human health have received considerable attention [1]. Accordingly, many experts in biomedical fields have analyzed the scientific articles accumulated in the literature database to assess the human health risk of EMF. Specifically, they classify the scientific articles according to the specific criteria, and then extract and summarize findings from the set of articles for an accurate assessment. However, the experts should manually review a massive number of scientific articles for classification. This process results in inefficiencies in terms of time and cost.

To address the problem, the use of neural network (NN) models is widely considered because the well-developed NN models can classify scientific articles with high accuracy. One of the main factors to determine the accuracy of the NN models is the word embeddings that convert the words into numerical vectors. However, the amount of available text data on the human health risk of EMF is not enough to train word embedding models. Therefore, it is necessary to select the appropriate pre-trained biomedical word embedding to develop the NN models for the classification of scientific articles on the human health risk of EMF.

In this paper, we conduct a comparative study on pre-trained word embeddings for the classification of scientific articles on the human health risk of EMF. First, the various NN models are trained using the three pre-trained biomedical word embeddings: PubMed-word2vec, BioWordVec, and PubMed-BERT [2–4]. Then, the performances of the trained NN models are evaluated to compare the pre-trained word embeddings. The results of the performance evaluation show that the NN models using PubMed-BERT outperform other NN models using PubMed-word2vec and BioWordVec in the classification of scientific articles on the human health risk of EMF. In particular, the CNN model using PubMed-BERT exhibits the highest performance by achieving an accuracy of 97.20%.

II. Dataset and Models

A. Dataset

To perform the comparative study, we created the dataset by collecting the article information from EMF-portal, which provides a list of 35,044 EMF-related articles and curated summaries of 6,943 articles. We first collected the titles of the scientific articles on the human health risk of EMF from the EMF-portal. Then, the abstracts of the articles were collected from PubMed, an open biomedical academic database. Finally, the label information of the articles was extracted from the curated summaries of EMF-portal and stored in the dataset. As a result, the dataset includes 3,362 articles on the human health risk of EMF with label information (i.e., 4 types of research category). Table I shows the summary of the dataset.

B. Models

Convolutional neural network (CNN), bi-directional long and short-term memory (BiLSTM), and simple NN are used to compare the pre-defined word embeddings. Fig. 1 shows the structures of the NN models. The CNN model includes the convolutional layer using 256 filters, the pooling layer, and the dense layer. The BiLSTM model includes the two BiLSTM layers having 64 units and 32 units, respectively, and the dense layer. Simple NN model includes one dense layer. Table II shows the hyper-parameters of the NN models.

III. Pre-trained word embeddings

In this section, we briefly introduce the pre-trained biomedical word embeddings that are compared in this work.

* ***Corresponding Author***

TABLE I. SUMMARY OF DATASET

Label	Number of articles	Ratio
Animal exposure experiment	1,196	35.57%
Cell exposure experiment	771	22.93%
Human exposure experiment	434	12.90%
Epidemiological study	961	28.58%
Total	3,362	-

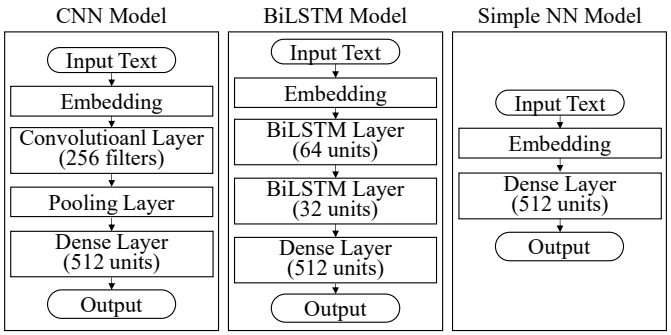

Fig. 1. Structures of NN models

TABLE II. HYPER-PARAMETERS OF NN MODELS

Hyper-parameters	Value
Input size (PubMed-word2vec)	650
Input size (BioWordVec)	650
Input size (PubMed-BERT)	200
Batch size	5
Loss Function	Cross entropy
Dropout	0.4
Optimization	Adam

TABLE III. PERFORMANCE COMPARISON OF NN MODELS USING DIFFERENT PRE-TRAINED WORD EMBEDDINGS

Embedding	Model	Precision (macro)	Recall (macro)	F1 (macro)	Accuracy (%)
PubMed-word2vec	CNN	0.954	0.954	0.953	96.16
	BiLSTM	0.906	0.908	0.903	91.92
	Simple NN	0.863	0.80.1	0.813	85.58
BioWordVec	CNN	0.959	0.951	0.954	96.42
	BiLSTM	0.936	0.929	0.927	94.46
	Simple NN	0.879	0.813	0.834	87.22
PubMed-BERT	CNN	**0.968**	**0.969**	**0.968**	**97.20**
	BiLSTM	0.940	0.938	0.933	94.26
	Simple NN	0.914	0.917	0.917	92.38

PubMed-word2vec includes 200-dimensional word embeddings for 4,087,446 biomedical words. The word embeddings of PubMed-word2vec are obtained by training the word2vec model with the title and abstract of articles on PubMed and PubMed central. BioWordVec is the open set of biomedical word embeddings including 200-dimensional word embeddings for 2,324,849 words. The word embeddings of BioWordVec are obtained by training the FastText model with text data on Medical Subject Headings (MeSH). Both PubMed-word2vec and BioWordVec provide word embeddings in the form of key-values where each word is mapped with one embedding vector.

PubMed-BERT is the pre-trained BERT language model, which is trained with the title and abstract of articles on MEDLIN and PubMed. PubMed-BERT does not provide word embeddings in the form of key-values. Instead, it takes the sentences as an input and generates the 768-dimensional embedding vectors as many as the user-defined number (126 by default). As the BERT model considers the start and end of input sentences, the generated word embeddings are context-aware.

IV. PERFORMANCE EVALUATION

For the precise comparison, we evaluated the performance of the NN models through K-fold cross-validation. In the K-fold cross-validation, the dataset was evenly divided into 5-folds (i.e., five sub-datasets). Then, the five experiments were conducted using four folds as training data and one fold as test data. Finally, the average values of the five experiments were used as the performance metric of the NN models.

Table III shows the performance comparison of the NN models using the different pre-trained word embeddings. In Table III, the NN models using PubMed-BERT exhibit better performance compared to the same type of NN models using other word embeddings. The PubMed-BERT generates the context-aware word embeddings. Thus, the NN models using PubMed-BERT can learn the more specific meaning of the input text and thereby outperform the other NN models using PubMed-word2vec and BioWordVec.

Furthermore, the CNN models achieve a higher performance compared to the other NN models using the same word embeddings. These results support that CNN-based models can achieve better performance in the classification of scientific articles on the human health risk of EMF compared to LSTM-based models and simple NN model.

V. CONCLUSION

In this paper, we compared the three pre-trained biomedical word embeddings (i.e., PubMed-word2vec, BioWordVec, and PubMed-BERT) for the classification of scientific articles on the human health risk of EMF. First, the CNN, BiLSTM, and simple NN models were trained using the three pre-trained word embeddings. Then, the performances of the NN models were evaluated. The result of the performance evaluation showed that the NN models using PubMed-BERT outperformed other NN models using PubMed-word2vec and BioWordVec in the classification of scientific articles on the human health risk of EMF.

ACKNOWLEDGMENT

This work was supported by the ICT R&D program of MSIT/IITP [2019-0-00102, A Study on Public Health and Safety in a Complex EMF Environment].

REFERENCES

[1] A.-K. Lee, S.-B. Jeon, and H.-D. Choi, "EMF Levels in 5G New Radio Environment in Seoul, Korea," IEEE Access, vol. 9, pp. 19716–19722, January 2021.

[2] S. V. Landeghem, F. Ginter, Y. V. Peer, and T. Salakoski, "EVEX: A PubMed-Scale Resource for Homology-Based Generalization of Text Mining Predictions," *in Proc. ACL-HLT 2011*, pp. 28–37, 2011.

[3] Y. Zhang, Q. Chen, Z. Yang, H. Lin, and Z. Lu, "BioWordVec, improving biomedical word embeddings with subword information and MeSH," Scientific Data, vol. 6, pp.1–9, May 2019.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Touatanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *in Proc. NAACL-HLT 2019*, pp. 4171–4186, 2019.