International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 / https://doi.org/10.14445/22315381/IJETT-V70I10P239

Volume 70 Issue 10, 402-407, October 2022
© 2022 Seventh Sense Research Group®

Original Article

Transformer Based Knowledge Graph Construction in Adverse Drug Reactions Prediction from Social Media Reviews

Arijit Dey[1], Jitendra Nath Shrivastava[2], Chandan Kumar[3]

[1]*Department of Computer Applications, B. P. Poddar Institute of Management and Technology, Kolkata, West Bengal, India.*
Department of Computer Science and Engineering, Invertis University, Bareilly, U. P., India.
[2,3]*Department of Computer Science and Engineering, Invertis University, Bareilly, U. P., India.*

[1] *Corresponding Author : ad.computerapplication@gmail.com*

Received: 01 August 2022 Revised: 13 October 2022 Accepted: 20 October 2022 Published: 23 October 2022

Abstract - *Adverse Drug Reaction (ADR) prediction is an essential research topic in the field of pharmacovigilance. It seeks the attention of many researchers nowadays. Reporting of ADR in social media is increasing as patients can directly share their opinions through various online forums, blogs, Twitter, etc. Several deep neural network models have been introduced to detect the presence of ADR in the data collected from social media and sometimes attain a promising outcome while sometimes not up to the mark. Recently knowledge graph embedding was introduced in the prediction of ADR with the most familiar Word2Vec approach in Natural Language Processing (NLP). However, Word2Vec suffers from converting large textual data and the context of data as this approach uses two common methods: Continuous Bag of Words and Skip-gram. This article proposes a new Transformer-Based Knowledge Graph to overwhelm the difficulties of a large corpus and the manifoldness of the words. This proposed model deals with building a knowledge graph from the token found in transformer models and passes it to the binary classifier to predict the presence or absence of ADR in a drug with an accuracy of 91%.*

Keywords *- Adverse Drug Reaction (ADR), Transformer, Knowledge Graph, Word2Vec, Natural Language Processing (NLP).*

1. Introduction

Adverse Drug Reactions (ADRs) are undesired outcome during the use of medicines that causes severe health problems and even cause death [11]. The clinical reviews of ADRs lead to a major burden in modern drug discovery [3]. Approximately 10% of European patients experienced the effects of ADRs [7]. Drugs are tested on animals and a group of people to identify ADRs before it is being marketed. But sometimes, it isn't easy to find the ADRs at an early stage of drug discovery [33]. Many researchers have applied different machine learning and deep learning methods to detect ADRs caused by a drug or a combination of drugs. These techniques also reduce the cost of drug development [21]. Advanced natural language processing tools, machine learning, and deep learning techniques make it possible to extract ADRs from biomedical literature in social media automatically. Nowadays, the researcher also pays attention to social media automatically extracting ADRs. Drug users share their views on several social media platforms like Twitter, Facebook, etc. Many researchers have given the importance of social media in the field of pharmacovigilence. They use Twitter data and many online medical forums as a source of their study [28, 30]. A review on extracting ADRs from social media has been done by [18]. They have shown several ways of social media approaches to identify and extract ADRs. They have used computerized methods to show the evaluation process in their study. They have evaluated their result based on precision (9/13, 69%), recall (9/13,69%), f-measure (6/13, 46%), accuracy (3/13, 23%), both true and false-positive rates (1/13, 8%), log-likelihood ratios (1/13,8%), support (1/13, 8%), confidence (1/13, 8%), leverage (1/13,8%), and Bayesian confidence propagation neural network (BCPNN) scores and variance (1/13, 8%). Zhang et al. [39] have proposed a novel method that can extract deep linguistic features and then combine them with shallow linguistic features for ADR detection, as they have reported that previous feature-based methods focus on extracting more shallow linguistic features that were unable to capture deep and subtle information in the context, ultimately failing to provide satisfactory accuracy. In today’s world, human knowledge integration is one of the new research paradigms of Artificial Intelligence. To solve complex problems, the ability to represent knowledge inspired by humans gives a new horizon in recent research [17, 25, 31]. A knowledge graph represents a structured form of facts. The relation between a pair of entities represents the properties in a welldefined manner in a knowledge graph [15]. A knowledge graph embedding model was introduced earlier in the prediction of ADRs. Zhang et.al.[37], have proposed a new knowledge graph embedding method based on the Word2Vec model in Nature Language Processing that embeds drugs and ADRs into their respective vectors and builds a logistic regression classification model to predict whether a given drug will have ADRs. Albadani et.al.[2], have introduced a transformer-based graph convolutional network to build a deep learning technique to predict the sentiment of a text. Due to the rapid growth of social media, it is very much needed to process the contextual data and mine the perspective of the data [12].

2. Related Work

The ADR reporting from a clinical trial has a major drawback; many ADRs can not be reported in due time. In this regard, social media like Twitter, Facebook, and several online forums play an important role in reporting ADRs directly by the patient or the drug user. A drug user or patient can directly share their opinion on a particular drug or group of drugs online, which could help the pharmacovigilance organization. So, NLP techniques must be developed to deploy machine learning and deep learning techniques. Many researchers proposed different NLP techniques and deployed several machine learning and deep learning approaches to predict ADRs from the review, which the patient shares online. A survey has been made on the different machine learning and deep learning techniques applied to different datasets and has been reported in an article by [4]. The researcher has applied knowledge graph embedding on sider and drugbank databases [37]. They have constructed KG embedding based on Word2Vec, a classical NLP method. Harnoune et al. [13] reported bert based knowledge graph for clinical data that defines the relationships between the biomedical entities. They have achieved high accuracy from their proposed framework. Huang [16] et al. have constructed predictive models for ADR detection from post-market surveillance and patients' posts on social media. They have concluded and reported in their research that deep learning NLP techniques outperform the traditional machine learning techniques in sentiment analysis nowadays after analysing various techniques. The bidirectional LSTM techniques were applied by Cocos [9] et al. to identify the side effects of drugs in their study on a sample Twitter dataset. Several deep-learning approaches have been applied to biomedical pieces of literature to detect ADRs [42]. Lee [41] et al. proposed hybrid deep learning techniques consisting of GCNN and BiLSTM for the prediction of side effects of drugs. Deep learning shows a new direction in healthcare research. CNN is applied in the recognition of patterns in the classification of data [14, 27]. Recent days knowledge graph shows a wide range of new directions in the field of pharmaceutical research [1, 36], biomedical research [23, 29, 34], and recommender system [35, 38]. Knowledge graph in biomedical research is very early [26]. The biomedical entities and the relationship between them have been extracted by a neural model reported by Li [20] et al. Challenges and research opportunities have been reported in [8]. Lordon [40] et al. have done a scoping review on social media data extraction of ADR.

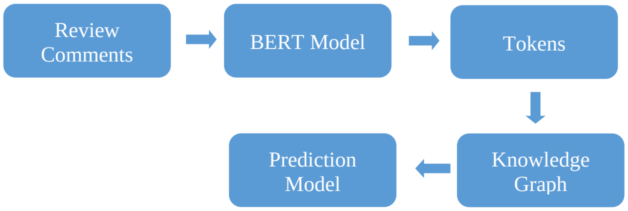

Fig. 1 The outline of the proposed methodology

3. Proposed Methodology

This article works on the reviews collected from patients. This section describes the outline of the proposed methodology, and the flowchart of the proposed methodology is shown in Figures 1 and 2. The first step of the proposed framework describes data collected from the data source; in the next step, some preprocessing is performed over the data found, passes through the BERT encoder system for tokenization, and finds the entities and relationships between a pair of entities. Then it introduces the knowledge graph construction. Then, the vectorization method was applied to entities and relationships to pass to the final step. The final step introduces a binary classifier to predict a drug's presence or absence of ADR. The Entire process workflow of the proposed methodology is shown in Figure 3.

3.1. Data Collection

This section describes the raw data collected from the data source [22]. After Searching several data sources, this data set has been chosen for getting the patient’s opinion on a drug. As mentioned above, knowledge graph embedding was done earlier on the data repository insider and drugbank. But no such work has been done till today on the review dataset, which directly depends on the patient’s opinion.

3.2 Data Pre-Processing

Some data pre-processing is needed as the patient reviews contain punctuation, mentions, numbers, etc. The original dataset received reviews along with ratings ranging from 1 to 10. The higher order rating mentioned no presence of ADR, and the lower order rating suggested a chance of ADR presence. This section introduces a technique by which rating 9 and 10 is converted to 0 means no ADR is present, and the rest of the ratings are converted to 1, which means the presence of ADR, as this article is working only on the presence and absence of ADRs in a drug.

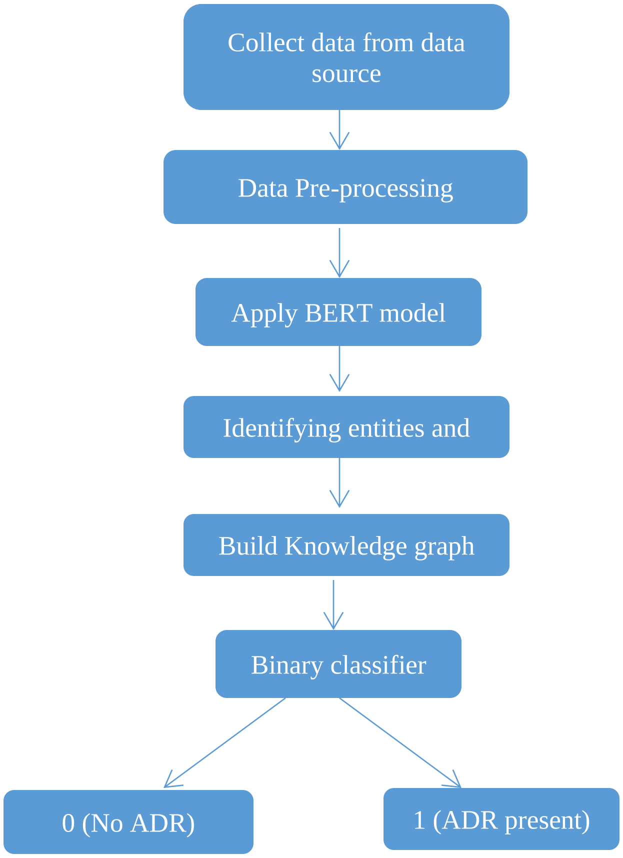

Fig. 2 Flowchart of the proposed methodology

3.3. BERT Encoder

This section describes a well-known and straightforward algorithm for feature extraction with text information. Words are treated as features in NLP techniques. Word2Vec is used widely in feature selection using NLP techniques [24]. Word2Vec is one hot encoding representation, and it uses two methods CBOW (common bag of words and skip-gram). The out-of-vocabulary words can not be handled by the Word2Vec method. This article presents a deep learning-based approach that proposes the BERT (Bidirectional Encoder Representations from Transformers) model. Here BERT has been used to understand the text language, and fine-tuning Bert has been introduced to build NLP for identifying entities and their relationships [6, 10]. The problem of vanishing gradient for long text processing in neural network models is handled with an attention mechanism. The neural network model converts the long text into a fixed-size vector. At the same time, the attention mechanism uses the decoder to retrieve the most significant context of the text [5].

3.4. Knowledge Graph

A knowledge graph is a process of storing some data that results from the information or text-based task. It puts our data into context and tracks all the data that links to machine learning. A knowledge graph is a kind of semantic network with added data. Knowledge graphs and machine learning can systematically improve accuracy performance [37]. The knowledge graph needs the entities and relationships between them. The text document is split into sentences to build a knowledge graph and assign context token vectors. A new knowledge graph embedding method is used to represent drug text reviews. This experiment shows that using a knowledge graph is very effective, which may make the scope of the gathered information by featuring entities. The nodes represent the entities, and the edges represent the relationship between a pair of entities. In this article, the triple is defined as (drug name, has, side effect) where drug name and side effect are two entities and describes the relationship between them.

3.5. Prediction Model

After the knowledge graph construction, predicting whether a drug contains ADR is required. The article will predict the side effects associated with a drug from the triple (drug name, has, side effect). Therefore, this triple defines a binary classification problem to identify a drug's presence or absence of ADR. This paper uses a support vector classifier to classify a drug's presence or absence of ADR. The support vector classifier draws two clear margins parallel to a hyperplane, which separates two classes in a high-dimensional space. Here '1' represents the presence and '0' represents the absence of ADR in a drug described in the following equation 1:

$$\boldsymbol{y} = \begin{cases} \mathbf{1}, & \boldsymbol{presence\ of\ ADR\ in\ drug} \\ \mathbf{0}, & \boldsymbol{absence\ of\ ADR\ in\ drug} \end{cases} \qquad (\mathbf{1})$$

4. Result Analysis

The performance of the proposed model is measured through the evaluation criteria accuracy and f1 score. The accuracy is measured by equation 2, and to find the f1 score, it is required to calculate precision and recall. Equation 3 computes the precision, and equation 4 computes the recall. Finally, f1 is calculated by equation 5. Figure 3 shows the ROC curve for the drug classifier. TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative, respectively. Table 1 shows the final prediction result accuracy got using the support vector classifier.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \qquad (2)$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (3)$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (4)$$

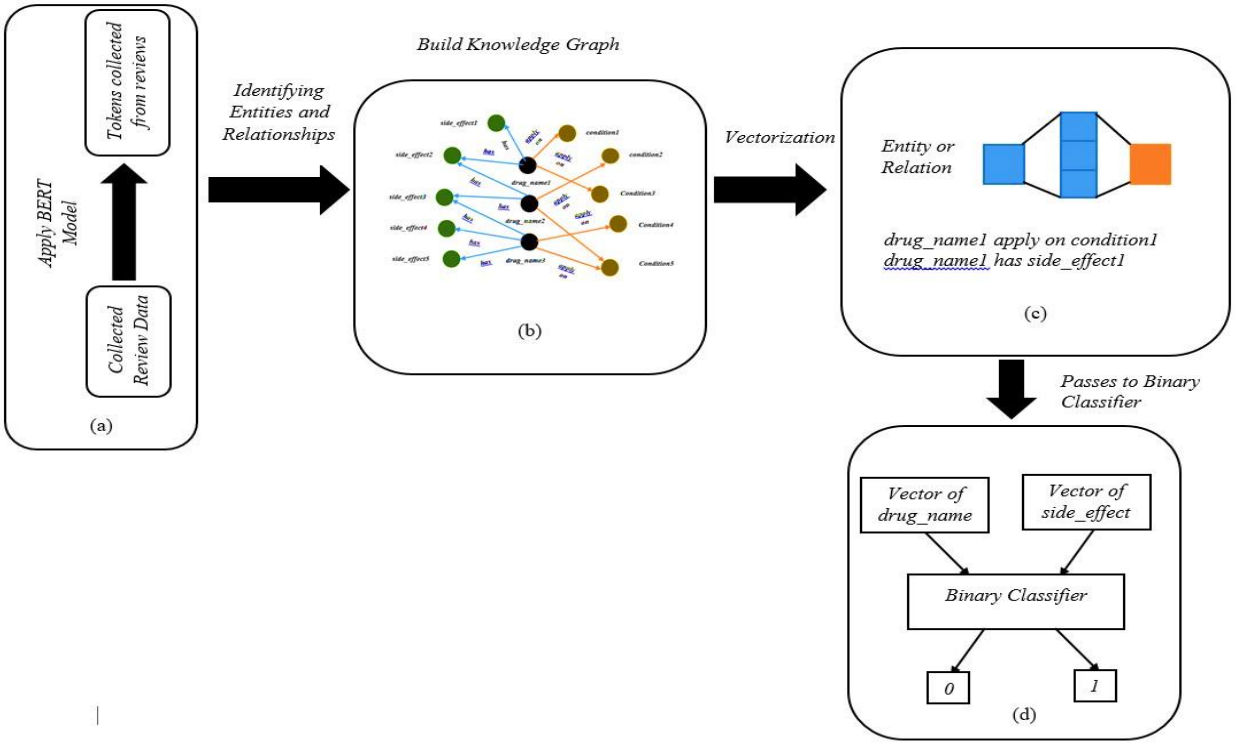

Fig. 3 Entire process workflow of proposed methodology (a) Review text tokenized through BERT model, (b) Build knowledge Graph from Entities and Relationships, (c) Vectorization process, (d) Binary Classifier for classifying presence and absence of ADR in a drug.

$$\mathrm{f1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \qquad (\mathbf{5})$$

Table 1. Accuracy Measurement

	precision	recall	f1 – score
0	0.90	0.91	0.91
1	0.91	0.91	0.91
accuracy		0.91	

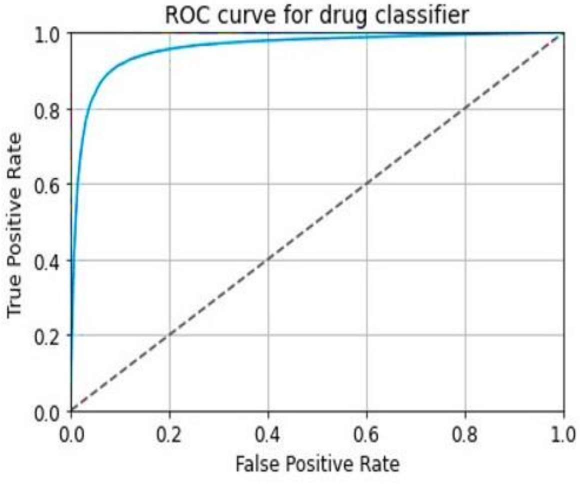

Fig. 4 ROC curve for drug classifier

5. Conclusion

This research proposes a Transformer-Based Knowledge Graph for predicting a drug's presence or absence of ADR. The article introduces the problem as a semantic representation of a text which extracts the knowledge representation of the text in a different context. For example, the following text is input text: “the bank is situated in the river bank.” Here, the word “bank” presents twice in the sentence with two different meanings or contexts. Bag-of words technique has the same vector as the word ’ bank.’ But it does not seek any attention to the context of words, while the encoder-based transformer technique can have the ability to find the context of words in which it has been used. This transformer-based knowledge graph finds the context of each word in the sentence and then gathers knowledge to identify the entities and relationships between a pair of entities. So, the triple of the knowledge graph can easily find out and represented as a semantic net. Finally, Vectorization is applied to the entities and relationships to go through to the binary classifier. This article uses a support vector machine to classify a drug's presence and absence of ADR. Searching is a very tedious job from a large pool of data. This proposed model makes this job easier and helps doctors and patients. The proper contextual information regarding the presence or absence of ADRs in a drug or a combination of drugs is found in this proposed model. The pharmaceutical organization should also benefit from this proposed model's functionalities.

References

[1] Abdelaziz I, Fokoue A, Hassanzadeh O, Zhang P, and Sadoghi M, “Large-Scale Structural and Textual Similarity-Based Mining of Knowledge Graph to Predict Drug–Drug Interactions,” *Journal of Web Semantics,* vol. 44, pp. 104-117, 2017.

[2] AlBadani B, Shi R, Dong J, Al-Sabri R, and Moctard O. B, “Transformer-Based Graph Convolutional Network for Sentiment Analysis,” *Applied Sciences*, vol. 12, no. 3, pp. 1316, 2022.

[3] Allison M, “Reinventing Clinical Trials,” *Nature Biotechnology,* vol. 30, no. 1, pp. 41-49, 2012.

[4] Arijit D, J., N., S., Chandan K, and Subhadip C, “Adverse Drug Reactions Extraction from Social Media: A Systematic Review,” *Grenze ID: 01.GIJET.8.1.11,* 2022.

[5] Bahdanau D, Cho K, and Bengio Y, “Neural Machine Translation by Jointly Learning to Align and Translate,” arXiv preprint arXiv:1409.0473, 2014.

[6] Biseda B, and Mo K, “Enhancing Pharmacovigilance with Drug Reviews and Social Media,” arXiv preprint arXiv:2004.08731, 2020.

[7] Bouvy J. C, De Bruin M. L, and Koopmanschap M. A, “Epidemiology of Adverse Drug Reactions in Europe: A Review of Recent Observational Studies,” *Drug Safety,* vol. 38, no. 5, pp. 437-453, 2015.

[8] Chen X, Jia S, and Xiang Y, “A review: Knowledge Reasoning Over Knowledge Graph,” *Expert Systems with Applications,* vol. 141 pp. 112948, 2020.

[9] Cocos A, Fiks A. G, and Masino A. J, “Deep Learning for Pharmacovigilance: Recurrent Neural Network Architectures for Labeling Adverse Drug Reactions in Twitter Posts,” *Journal of the American Medical Informatics Association,* vol. 24, no. 4, pp. 813-821, 2017.

[10] Devlin J, Chang M.W, Lee K, and Toutanova K, “Bert: Pre-training of Deep Bidirectional Trans- Formers for Language Understanding,” arXiv preprint arXiv:1810.04805, 2018.

[11] Edwards I. R, and Aronson J. K, “Adverse Drug Reactions: Definitions, Diagnosis, and Management,” *The Lancet,* vol. 356, no. 9237, pp. 1255-1259, 2000.

[12] Habimana O, Li Y, Li R, Gu X, and Yu G, “Sentiment Analysis using Deep Learning Approaches: An Overview,” *Science China Information Sciences,* vol. 63, no. 1, pp. 1-36, 2020.

[13] Harnoune A, Rhanoui M, Mikram M, Yousfi S, Elkaimbillah Z, and El Asri B, “Bert Based Clinical Knowledge Extraction for Biomedical Knowledge Graph Construction and Analysis,” *Computer Methods and Programs in Biomedicine Update*, vol. 1, pp. 100042, 2021.

[14] Hirohara M, Saito Y, Koda Y, Sato K, and Sakakibara Y, “Convolutional Neural Network Based on Smiles Representation of Compounds for Detecting Chemical Motif,” *BMC Bioinformatics,* vol. 19, no. 19, pp. 83-94, 2018.

[15] Hogan A, Blomqvist E, Cochez M, d’Amato C, Melo G. D, Gutierrez C, Kirrane S, Gayo J. E. L, Navigli R, Neumaier S, et al., “Knowledge Graphs,” *Synthesis Lectures on Data, Semantics, and Knowledge,* vol. 12, no. 2, pp. 1-257, 2021.

[16] Huang J.Y, Lee W.P, and Lee K.D, “Predicting Adverse Drug Reactions from Social Media Posts: Data Balance, Feature Selection and Deep Learning,” *In Healthcare,* MDPI , vol. 10, pp. 618, 2022.

[17] Ji S, Pan S, Cambria E, Marttinen P, and Philip S. Y, “A Survey on Knowledge Graphs: Representation, Acquisition, and Applications,” *IEEE Transactions on Neural Networks and Learning Systems,* vol. 33, no. 2, pp. 494-514, 2021.

[18] Lardon J, Abdellaoui R, Bellet F, Asfari H, Souvignet J, Texier N, Jaulent M.C, Beyens M.N, Burgun A, Bousquet C, et al., “Adverse Drug Reaction Identification and Extraction in Social Media: A Scoping Review,” *Journal of Medical Internet Research,* vol. 17, no. 7, pp. e4304, 2015.

[19] Rotmensch M, Halpern Y, Tlimat A, Horng S, and Sontag D, “Learning a Health Knowledge Graph from Electronic Medical Records,” *Scientific Reports,* vol. 7, no. 1, pp. 1-11, 2017.

[20] Li F, Zhang M, Fu G, and Ji D, “A Neural Joint Model for Entity and Relation Extraction from Biomedical Text,” *BMC Bioinformatics,* vol. 18, no. 1, pp. 1-11, 2017.

[21] Li J, Zheng S, Chen B, Butte A. J, Swamidass S. J, and Lu Z, “A Survey of Current Trends in Computational Drug Repositioning,” *Briefings in Bioinformatics,* vol. 17, no. 1, pp. 2-12, 2016.

[22] Michael Allen, “1804 Python Healthcare,” 2020.

[23] Moon C, Jin C, Dong X, Abrar S, Zheng W, Chirkova R. Y, and Tropsha A, “Learning Drug-Disease-Target Embedding (DDTE) from Knowledge Graphs to Inform Drug Repurposing Hypotheses,” *Journal of Biomedical Informatics,* vol. 119, pp. 103838, 2021.

[24] Naseem U, Razzak I, Khan S. K, and Prasad M, “A Comprehensive Survey on Word Representation Models: From Classical to State-of-the-Art Word Representation Language Models,” *Transactions on Asian and Low-Resource Language Information Processing,* vol. 20, no. 5, pp. 1-35, 2021.

[25] Newell A, Shaw J. C, and Simon H. A, “Report on a General Problem Solving Program,” *In IFIP Congress,* Pittsburgh, PA, vol. 256, pp. 64, 1959.

[26] Nicholson D. N, and Greene C. S, “Constructing Knowledge Graphs and their Biomedical Applications,” *Computational and Structural Biotechnology Journal,* vol. 18, pp. 1414-1428, 2020.

[27] O¨ ztu¨rk H, O¨ zgu¨r A, and Ozkirimli E, “Deepdta: Deep Drug-Target Binding Affinity Prediction,” *Bio-Informatics,* vol. 34, no. 17, pp. i821–i829, 2018.

[28] Patki A, Sarker A, Pimpalkhute P, Nikfarjam A, Ginn R, O'Connor K, Smith K, and Gonzalez G, "Mining Adverse Drug Reaction Signals from Social Media: Going Beyond Extraction," *Proceedings of BioLinkSig*, vol. 2014, pp. 1-8, 2014.

[29] Rotmensch M, Halpern Y, Tlimat A, Horng S, and Sontag D, "Learning a Health Knowledge Graph from Electronic Medical Records," *Scientific Reports,* vol. 7, no. 1, pp. 1-11, 2017.

[30] Sarker A, and Gonzalez G, "Portable Automatic Text Classification for Adverse Drug Reaction Detection via Multi-Corpus Training," *Journal of Biomedical Informatics,* vol. 53, pp. 196-207, 2015.

[31] Shortliffe E, "Computer-Based Medical Consultations: MYCIN," *Elsevier*, vol. 2, 2012.

[32] Mauik Panchal, Prof. Rutika Ghariya, "A Review On Detection of Fake News Using Various Techniques," *SSRG International Journal of Computer Science and Engineering*, vol. 8, no. 6, pp. 1-4, 2021. Crossref, https://doi.org/10.14445/23488387/IJCSE-V8I6P101.

[33] Whitebread S, Hamon J, Bojanic D, and Urban L, "Keynote Review: In Vitro Safety Pharmacology Profiling: An Essential Tool for Successful Drug Development," Drug Discovery Today, vol. 10, no. 21, pp. 1421-1433, 2005.

[34] Xu J, Kim S, Song M, Jeong M, Kim D, Kang J, Rousseau J. F, Li X, Xu W, Torvik V. I, et al., "Building a Pubmed Knowledge Graph," *Scientific Data,* vol. 7, no. 1, pp. 1-15, 2020.

[35] Yang Z, and Dong S, "Hagerec: Hierarchical Attention Graph Convolutional Network Incorporating Knowledge Graph for Explainable Recommendation," *Knowledge-Based Systems*, vol. 204, pp. 106194, 2020.

[36] Zeng X, Zhu S, Hou Y, Zhang P, Li L, Li J, Huang L. F, Lewis S. J, Nussinov R, and Cheng F, "Network-Based Prediction of Drug-Target Interactions using an Arbitrary-Order Proximity Embedded Deep Forest," *Bioinformatics*, vol. 36, no. 9, pp. 2805-2812, 2020.

[37] Zhang F, Sun B, Diao X, Zhao W, and Shu T, "Prediction of Adverse Drug Reactions Based on Knowledge Graph Embedding," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, pp. 1-11, 2021.

[38] Zhang S, Yao L, Sun A, and Tay Y, "Deep Learning Based Recommender System: A Survey and New Perspectives," *ACM Computing Surveys (CSUR),* vol. 52, no. 1, pp. 1-38, 2019.

[39] Zhang Y, Cui S, and Gao H, "Adverse Drug Reaction Detection on Social Media with Deep Linguistic Features," *Journal of Biomedical Informatics,* vol. 106, pp. 103437, 2020.

[40] Lardon J, Abdellaoui R, Bellet F, Asfari H, Souvignet J, Texier N, & Bousquet C, "Adverse Drug Reaction Identification and Extraction in Social Media: A Scoping Review," *Journal of Medical Internet Research*, vol. 17, no. 7, pp. e4304, 2015.

[41] Lee C. Y, and Chen Y.P. P, "Descriptive Prediction of Drug Side-Effects using a Hybrid Deep Learning Model," *International Journal of Intelligent Systems*, vol. 36, no. 6, pp. 2491-2510, 2021.

[42] Wang C.S, Lin P.J, Cheng C.L, Tai S.H, Yang Y.H. K, Chiang J.H., et al., "Detecting Potential Adverse Drug Reactions using a Deep Neural Network Model," *Journal of Medical Internet Research,* vol. 21, no. 2, pp. e11016, 2019.