Open Information Extraction Methodology for a New Curated Biomedical Literature Dataset

Nesma Abdel Aziz Hassan[1]*, Rania Ahmed Abdel Azeem Abul Seoud[2], Dina Ahmed Salem[3]
Department of Computer-Faculty of Engineering, Misr University for Science and Technology, Egypt[1]
Department of Electrical Engineering-Faculty of Engineering, Fayoum University Egypt, Fayoum[2]
Department of Computer-Faculty of Engineering, Misr University for Science and Technology, Egypt[3]

***Abstract*—The research articles contain a wealth of information about the interactions between biomedical entities. However, manual relation extraction processing from the literature by domain experts takes a lot of time and money. In addition, it is often prohibitively expensive and labor-intensive, especially in biomedicine where domain knowledge is required. For this reason, computer strategies that can use unlabeled data to lessen the load of manual annotation are of great relevance in biomedical relation extraction. The present study solves relation extraction tasks in a completely unsupervised scenario. This article presents an unsupervised model for relation extraction between medical entities from PubMed abstracts, after filtration and preprocessing the abstracts. The verbs and relationship types are embedded in a vector space, and each verb is mapped to the relation type with the highest similarity score. The model achieves competitive performance compared to supervised systems on the evaluation using ChemProt and DDI datasets, with an F1-score of 85.8 and 88.5 respectively. These improved results demonstrate the effectiveness of extracting relations without the need for manual annotation or human intervention.**

***Keywords*—Relation extraction; BERT; open information extraction; biomedical literature; ChemProt; DDI**

I. Introduction

With more than 32 million citations of biomedical literature in PubMed [1], the medical field is facing a substantial expansion of data, posing challenges for biomedical researchers efficiently and automatically to extract information about specific biomedical entities such as genes, proteins, and diseases. Gathering labeled data required for training and creating models for natural language processing (NLP) is a time-consuming task because it requires human annotation and can take a lot of effort to complete. Gathering a sufficient amount of labeled data for NLP can be challenging since language is complex and diverse. Models require large amounts of labeled data to learn from to generalize well, and the larger the dataset, the more challenging it becomes to label it. Also, the quality of the labeled data is crucial for the accuracy of the NLP model. To ensure high-quality labeled data, it is necessary to have multiple annotators to check for consistency, which can increase the cost of the labeling process. Overall, gathering labeled data for NLP is a time-consuming and resource-intensive process, which can make it expensive and difficult. There is a necessity for quick and scalable discovery, extraction, and organization of the information contained within this data; an issue that raises the need to use information extraction techniques. However, the process of information extraction is one of the primary difficulties in NLP, and its application to a vast amount of data has some restrictions.

Relation extraction and open information extraction are both techniques used in information extraction, but they have distinct differences in their approaches and goals. Relation extraction focuses on identifying and extracting specific pre-defined relationships or connections between entities mentioned in the text. It aims to discover structured relationships with predefined categories. The output of relation extraction is typically a set of entity pairs with their associated relationship type.

Open Information Extraction (OIE) is a relation-free, open-domain paradigm that enables unsupervised information extraction. Its major goal is to create a triple relation with the elements <entitiy1> <Relation> <entity 2> from unstructured data without having to predefine the relationship between the two entities. The extracted tuples can be binary, ternary, or n-ary, depending on how many entities are involved in the relationship. In the biomedical domain, OpenIE has been used to extract relationships between entities such as genes, proteins, diseases, and drugs from the scientific literature. This method may result in scalable and fast performance [2]. In contrast to OIE, RE demands relation definition prior to extraction. Similar to OIE, a binary relation or a higher-order relation (n-ary) can be found in the extracted relation [3]. Fig. 1 summarizes the difference between OIE and RE.

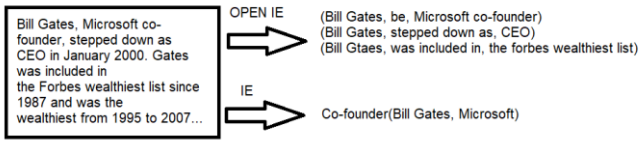

Fig. 1. Open information extraction vs. relation extraction.

Relation extraction is often used to populate or update a structured knowledge base with specific relationship types. The extracted relationships are typically mapped to existing entities in the knowledge base, providing structured information for further analysis.

While Open information extraction is useful for discovering new relationships or facts that may not be present in a pre-existing knowledge base, the extracted relational tuples can be integrated into a knowledge base to expand its coverage and provide additional information. In summary, relation

extraction focuses on extracting specific pre-defined relationships between entities, while open information extraction aims to extract as much information as possible without relying on predefined relationship types. Relation extraction is often supervised and used for structured knowledge base populations, whereas open information extraction is more flexible, unsupervised, and useful for discovering new information.

OIE is an essential NLP task, and because of its many potential uses in information retrieval, information extraction, text summarization, and question answering, it was chosen as a source task to transfer to other NLP tasks [2]. Even though several OIE algorithms have been created in the last ten years, only a few studies [3, 4] have attempted to address Unsupervised Relation Extraction (URE) using machine learning (ML) and deep learning methods. Due to the shortage of labeled data, researchers have recently been more and more interested in model generalization in deep learning. Fig. 2 shows the number of publications in Biomedical literature in the last years. It is shown that there is a steady increase in the number of publications in biomedical literature according to PubMed and Web of Science databases.

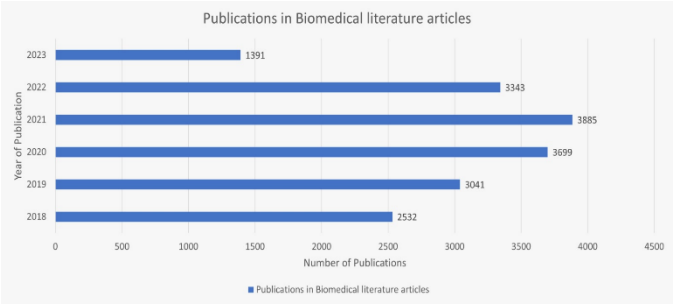

Fig. 2. Number of publications in biomedical literature.

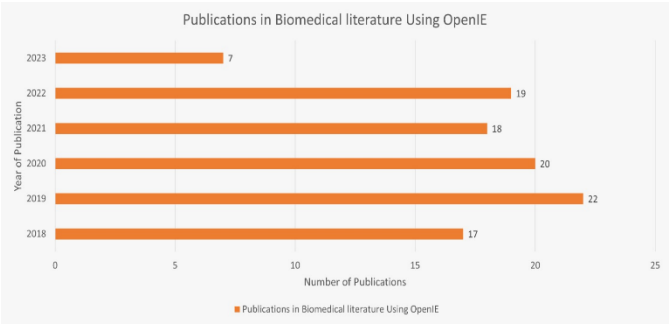

Fig. 3. Number of publications in biomedical literature using OpenIE.

Fig. 3 depicts a few articles that use OpenIE method. That's why there is a growing need for unsupervised models in biomedical relation extraction due to the exponential increase in biomedical literature and manual annotation of all relevant articles is becoming more and more difficult and expensive. OpenIE extracts relationships between medical entities from unstructured text, which can provide valuable insights into the underlying mechanisms of diseases and drug actions by the integration of domain-specific language models that are trained on biomedical literature data, it can capture the nuances of the language and improve the accuracy of the extracted relationships. Furthermore, recent advances in deep learning have enabled the development of more powerful language models, such as BERT [5] and GPT-3 [6], which can capture even more complex linguistic patterns and improve the accuracy of the extracted relationships.

Overall, the integration of domain-specific language models in OIE can improve the accuracy and efficiency of relation extraction from biomedical literature and enable researchers to discover new insights and relationships in the field.

This paper proposes an approach for Open Information Extraction of biomedical literature in a completely unsupervised scenario. This approach identifies interaction words using part-of-speech tagging (POS) in the subject-verb-object form to fully utilize both whole and partial phrase structure information. The interacted verb is then mapped to the nearest group of predefined relations, allowing the system to classify and perform relation extraction without human intervention or the need for labeled data. The resulting data is fed into a pre-trained model, specifically Bidirectional Encoder Representations from Transformers (BERT), which performs fine-tuning with the self-created dataset to enable classification and relation extraction.

The rest of this paper is organized as follows Section II surveys previous work on OIE, Section III explains the materials and methods used to conduct this research, and Section IV evaluates the proposed model by summarizing the results and discussing the findings. Finally, Section V concludes the paper.

II. Related Work

The amount of biomedical literature is growing rapidly, and manual annotation of all relevant articles is becoming more and more difficult and expensive. Additionally, supervised models may not generalize well to new domains and contexts, which limits their utility in real-world applications. OpenIE is particularly useful for extracting relations and facts from large and complex biomedical literature, which can provide valuable insights into the interactions between biological entities and the underlying mechanisms of various diseases [7]. This literature review aims to provide an overview of recent studies on the use of biomedical relation extraction, highlighting the state-of-the-art methodologies and potential future directions for research in this field.

Drug descriptions from Wikipedia and DrugBank were used by Zhu et al. [8] to provide semantic data about drug entities to the BERT model. To obtain sentence representation with entity information, mutual drug entity information, and drug entity information, they used three different types of entity-aware attentions. The BioBERT embeddings of two medications were subtracted to get the mutual information vector of the two drug entities. On the DDI corpus, they reported an 80.9 (micro F1-score).

By linking one or more natural language questions to each relation, Levy et al. [9] have reframed the relation extraction job as a reading comprehension issue. In a zero-shot context, this method enables generalization of undiscovered relations. You [10] proposed a Path-Based MiRNA-Disease Association (PBMDA) prediction model that can infer potential miRNA-

disease associations by combining known human miRNA-disease associations, miRNA functional similarity, disease semantic similarity, and Gaussian interaction profile kernel similarity for miRNAs and diseases.

Similarly, in a study by Zaikis et al. [11], the authors proposed an approach that integrates a prior knowledge by using BioBERT which outperforms existing state-of-the-art methods on the DDI corpus in both drug named entity recognition and overall DDI extraction task. Gu et al. [12] proposed a Convolutional Neural Network (CNN) model to learn a more robust relation representation based on both word sequences and dependency paths for the CID relation extraction task, which could naturally characterize the relations between chemical and disease entities.

Drug-drug interactions (DDIs) prediction and extraction model based on BioBERT was proposed by Mondal [13] and Zhu [14]. Both experiments on the DDIExtraction 2013 corpus demonstrate that it can outperform baseline architectures in F1-score, which is a good illustration of BioBERT's use in the field of biomedical text processing.

Peng et al. [15], used a distant supervision model to extract extraction of CIDs from biomedical literature. The authors proposed a method that combined novel statistical features with machine learning model to improve the accuracy of CID extraction. The results showed that their method achieved high precision and recall in identifying chemical-induced disease (CID) relations.

Chen [16] provided a computational approach to identify potential miRNA-disease connections that significantly cut down on experimental time and expense. They created the WBSMDA model for within and between Score for MiRNADisease Association prediction to predict miRNAs linked with diseases. Mario [17], two new relation extraction methods that were proposed to extract the mutation-disease relationship outperform conventional techniques.

In, Zhang et al. [18] also proposed a hybrid model based on RNNs and CNNs to classify PPIs and DDIs. The inputs of the hybrid model are sentence sequences and SDPs generated from the dependency graph. RNNs and CNNs models were employed to learn the feature representation from sentence sequences and SDPs, respectively.

Although the studies discussed above have made significant contributions to biomedical relation extraction, there are still limitations to consider. For instance, some studies rely on specific datasets, making them less generalizable to other domains or contexts. For example, the studies by Mondal [13] and Zhu [14] focused solely on drug-drug interactions and may not be applicable to other types of biomedical relations. Additionally, some studies may require significant domain-specific knowledge and preprocessing, such as the study by Peng et al. [15], which utilized distant supervision and statistical features. While the results of these studies are promising, the performance of the proposed models may be limited by the quality and representativeness of the training data, which can introduce biases and limit the generalizability of the models.

In summary, while most high-performing biomedical relation extraction systems to date are based on supervised approaches, this method requires large amounts of labeled data that can be expensive and time-consuming to obtain. Furthermore, supervised approaches may be biased and limited in generalizability to new domains or contexts. Alternatively, unsupervised models can learn directly from data and can be applied to new domains and contexts without the need for retraining on newly labeled data. The exponential increase in biomedical literature and the limitations of supervised models highlight the growing need for unsupervised models in biomedical relation extraction.

III. METHODOLOGY

By using this approach, the study aims to overcome the limitations of supervised methods and provide a scalable and efficient solution for relation extraction in the biomedical domain. The end-to-end framework described in Fig. 4 involves the following steps:

- Downloading and preprocessing records from PubMed using specific queries (each document contained the title and the abstract, PMID).
- Converting the data into sentence pair format for the Named Entity Recognition (NER) process.
- The medical entities such as genes, proteins, diseases, and drugs are captured in NER and replaced with predetermined tags.
- To extract the relations of the form entity1-verb-entity2, a part-of-speech (POS) tagger is used. By using the POS tags, the model can identify the verbs that connect the entities.
- Creating a list of relationship types that are of interest in the task at hand to map the extracted verbs to them. These relations would then be mapped to the verbs that are extracted using a part-of-speech tagger. Once the set of relations has been defined, the extracted verbs and relationship types are embedded in a vector space and each verb is mapped to the relation type that has the highest score after calculating the similarity score.
- The resulting data from the previous steps are used to finetune the pre-trained BERT model.

A. Material

The most significant databases for biomedical literature include MEDLINE, PubMed, Scopus, Embase, and Web of Science. The model at hand gathers literature primarily from the PubMed database. The PubMed database contains more than 35 million citations and abstracts from biomedical literature. In order to improve both individual and planetary health, PubMed promotes the free search and retrieval of biomedical and life sciences literature.

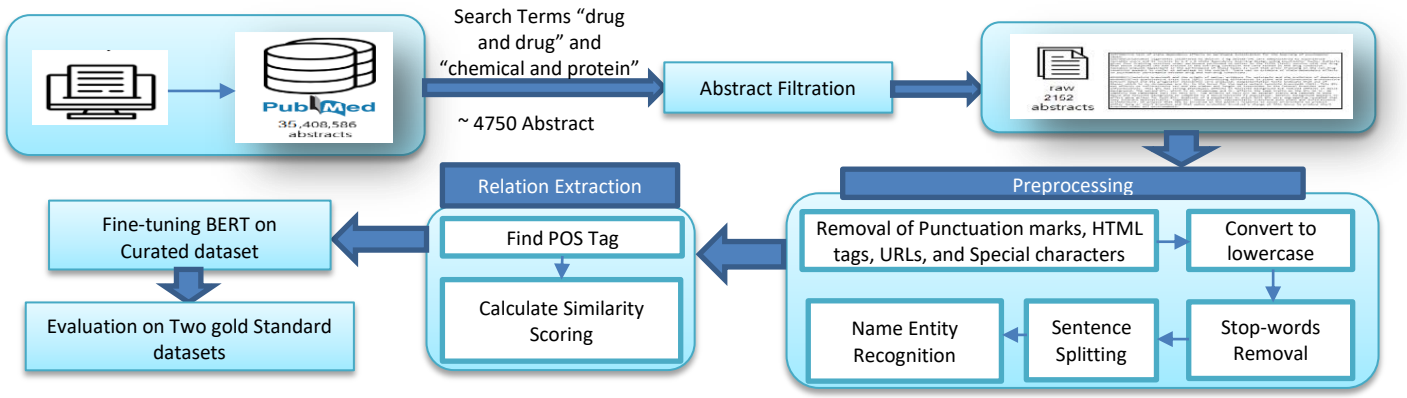

Fig. 4. An end-to-end framework for the proposed model.

While journal articles are not included in full text, PubMed [1] provides links to the full texts when they are available from other sources, such as the publisher's website or PubMed Central (PMC). Each article in PubMed is assigned a unique identification code called a PMID. PMIDs are never reused and do not change over time or throughout processing, providing a reliable way to track and reference articles in the database. PubMed is a valuable resource for researchers in the biomedical field, providing access to a vast amount of literature that can inform scientific discoveries, and advance healthcare. The online availability of PubMed to the public has also made it a valuable resource for patients and healthcare professionals seeking information on medical conditions and treatments [19].

B. Articles Selection and Generation of Training Data

Data must be carefully selected to support academic requirements, scientific research, and other purposes. Data curation is the process of locating, organizing, and managing data collections so that consumers looking for information can use them. Choosing an accurate corpus surely will have a positive impact on the study. The framework proposed in [19] for title retrieval can be expanded to extract abstracts as well. This involves using similar techniques to those used for title extraction but applied to the abstracts. A PubMed search using the terms "drug and drug" and "chemical and protein" was used as the starting point of the process. The two search queries resulted in 4750 abstracts. A sample of selected publications' titles, publication dates, and PMids are displayed in Table I.

TABLE I. SAMPLE OF SELECTED PAPERS AND THEIR DATE OF PUBLICATION

PMID	TITLE	Date of Publication
37311160	Breast Cancer Therapeutics and Hippo Signaling Pathway: Novel MicroRNA-Gene-Protein Interaction Networks.	2023
36387483	Assessment of potential drug-drug interactions among outpatients in a tertiary care hospital: focusing on the role of P-glycoprotein and CYP3A4 (retrospective observational study).	2022
34338261	Advances in chemical probing of protein O-GlcNAc glycosylation: structural role and molecular mechanisms.	2021
29927582	Chemical and Biochemical Perspectives of Protein Lysine Methylation.	2019

The filtration stage included removing all articles that are published before 2019, removing articles that don't have an abstract, removing duplicated articles (the same article appears in different queries) using the pmid. After applying these filters, the resulting dataset contained 1052 abstracts. The abstracts from the two search queries were merged to form one dataset. Fig. 5 summarizes the main steps of articles selection.

Fig. 5. Paper selection and dataset creation steps.

1) Preprocessing: Once the model has finished selecting relevant articles for the literature curation step, the preprocessing stage starts by splitting the paragraphs into sentences. This conversion allowed for easier identification of relations between two entities within a sentence rather than a paragraph. The extracted sentences were preprocessed before being used to train an algorithm and before a trained model was applied to them because biomedical texts may contain a wide variety of words, numbers, URLs and links, names of genes, chemicals, diseases, and so forth, as well as abbreviations which may be discussed, and all of these may add noise, rather than aid in the classification task. The next stage was to lowercase every word while removing any URLs, numerals, and punctuation from a phrase. The words "no," "not," "nor," "ae," "aes," and "adr" were left out, as well as any other words with three letters or less. A set of customized stop words was finally eliminated. We also performed additional preprocessing steps, including special characters, white spaces, and punctuation. These steps constituted the

extent of the preprocessing conducted in this phase. All the previous steps were done using the Scispacy library [20], which is tailored for biomedical, scientific, and clinical material and Natural Language Toolkit (NLTK) [21], which provides a range of efficient natural language algorithms.

The next step after preprocessing was Name Entity Recognition (NER) for extracting the medical entities. We utilized Scispacy to identify and extract the biomedical entities from the extracted phrases. Following NER, we filtered out the sentences from the abstracts. Sentences that did not contain any entities or contained only one entity were removed, while sentences with two entities were considered the target sentences for performing relation extraction and training the model. The total number of sentences with two entities or more was 3407 sentences. A total of 27,123 text descriptions are obtained by the proposed model which are labeled as gene, 19,471 are labeled as disease, 30,289 are labeled as chemical and 25,013 are labeled as protein from the NER task.

2) Relation extraction: A two-step algorithm was used to create a text classification model for the purpose of identifying Chemical-protein or drug-drug-related relations. The first step was part of speech tagging used to extract relations in the form of subject-verb-object. The second step was the resulting verbs tagged from the POS stage and the relation types that we are interested in learning are embedded in a vector space, and each verb is mapped to the relation type that it is most comparable to. To make this procedure fully automatic, it is necessary to specify the collection of relation types of interest and set a threshold for verb mapping, below which no relation class is assigned. Here's the process followed:

a) Part of speech tagging: Initially, Part of Speech (POS) tagging was employed to identify all the verbs present in the sentences not only the main verb between the medical entities. The POS tagger is a tool that assigns a part of speech (such as a noun, verb, adjective, or adverb) to each word in a sentence. By using the POS tags, the model can identify the verbs that connect the entities.

b) Similarity scoring: This step involves identifying the verbs in the sentence that are likely to indicate a relationship between the named entities of interest, such as genes, proteins, drugs, and diseases. The following verbs—increase, decrease, cause, treat, prevent, combine, reduce, and bind—were used to create a fixed set of relations. Once the set of relations has been defined, the extracted verbs and relationship types are embedded in a vector space. This involves representing the verbs and relationship types as vectors in a high-dimensional space, where the similarity between the vectors reflects the semantic similarity between the verbs and relationship types. Each verb is then mapped to the relation type that has the highest score after calculating the similarity score. This involves calculating the cosine similarity between the vector representation of the verb and the vector representation of each relation type and selecting the relation type that has the highest similarity score.

TABLE II. EXAMPLES OF VERB MAPPING AND CALCULATING SIMILARITY

Verb	Relation	Similarity
[‘Reduced’]	Decrease	0.553579
['recovered']	treat	0.535033
['entail', 'including', 'increased', 'eliminated']	increase	0.68346
['used', 'prevent']	Prevent	1
['impaired', 'blocked', 'leading']	cause	0.565928
['regulate', 'serve']	Bind	0.46995
['prepared']	No relation	0.166231

Table II shows a sample of extracted verbs, and their mapping to the set of relations. The similarity scoring threshold was set at 0.4. In cases where multiple verbs exceeded the threshold, the verb with the highest similarity score was selected. Verbs that achieved a similarity score below 0.4 were mapped to a "no relation" label. The RE step is crucial because it guarantees that the resulting relations will have high precision (although at the expense of low recall), as they substantially rule out the chance of the two entities randomly co-occurring in the sentence through the subject-verb-object relationship. In other words, we arrive at a limited, but highly valuable collection of relations that can be applied in a manner that is similar to distant supervision.

3) Transformer tuning of self-created dataset: Language models have demonstrated improved performance in various NLP tasks by considering contextual information when representing features. Popular language models like ELMO [22], BERT [5], and ULM Fit [23] are commonly used in NLP tasks.

a) Bidirectional Transformers (BERT): Among them, BERT introduced by Google in 2019, has gained immense popularity for its ability to pre-train deep bidirectional representations on unlabeled text. By considering context from both sides in all 12 transformer layers, BERT has shown enhanced performance in many NLP applications, including relation extraction.

BERT was used to train the RE classifier, it was fine-tuned for up to 5 epochs because early experiments revealed that the model began overfitting to noise and that the validation loss increased beyond that. The typical BERT pre-processing method for relation extraction is used to replace the entity names in the phrase with predetermined tags so that the BERT model can recognize the entities in the sentence. All references to proteins, drugs, and chemicals are expressly changed to @PROTEIN$, @DRUG$, and @CHEMICAL$, respectively. Fig. 6 shows an example of sentences after preprocessing for BERT.

Sentence examples
@DRUG$ decreases the elimination of @DRUG$ causing an increase in overall exposure.
Anticoagulants @DRUG$ may increase sensitivity to oral @DRUG$
@CHEMICAL$ potently attenuated gene expressions involved in inflammation, such as iNOS, COX-2 and @GENE$.
They suggest that TRPML1 works in concert with @CHEMICAL$ to regulate @GENE$ translocation between the cytoplasm and lysosomes.

Fig. 6. Examples of sentences preprocessing before feeding them to BERT.

BERT uses WordPiece embedding [24] to address the out-of-vocabulary issue, and each input word will be divided into subwords from the WordPiece vocabulary. Additionally, we add the segment and position information of the tokens in the input phrase using segment embedding and position embedding, both of which follow the original BERT input structure.

When using BERT, it is necessary to select a suitable model and adjust hyperparameters such as learning rate, batch size, and the number of epochs. The BERT-base uncased model is a common choice for sequence classification, it was found that the uncased version performed better [25,26,27,28]. During training, a trained model is validated to assess overfitting and estimate performance before applying it to the target application. We found that validating specific combinations of claim and description concatenation parts, which are not specifically trained but used in other combinations, has low significance.

b) Hyperparameter settings: Hyperparameter tuning is an important step in training machine learning models to achieve optimal performance. In this study, the Transformers library was used to perform hyperparameter tuning for the BERT-Base-Uncased model used in biomedical relation extraction.

The experiments involved varying different hyperparameters, including batch size, learning rate, maximum sequence length, and number of epochs. The maximum sequence length varied between 128 and 512, the batch size varied between 4 and 16 and the learning rate varied between 1e-5 and 4e-5. The results showed that the model achieved minimum loss when the learning rate was around 1e^-5. Fig. 7 shows the relationship between loss and the learning rate. The maximum success value was reached with a batch size fixed at 8, and a maximum sequence length equal to 256. Table III represents the ideal parameter values after experimenting with various parameter value combinations.

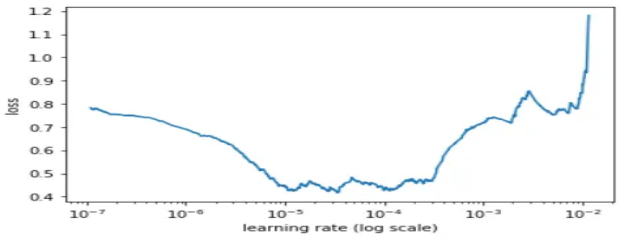

Fig. 7. Plotting of loss of the model versus learning rate.

TABLE III. HYPERPARAMETER SETTINGS FOR BERT

Hyperparameter	*BERT fine-tuning*
Epochs	3,5
learning rate	$1e^{-5}$
Optimizer	Adam
Batch Size	8
Max Sequence length	256
Early stopping	NO

Using the Adam optimizer, the BERT-Base-Uncased model was improved. The best-performing models are saved and evaluated on the test set using the epoch with the highest Recall score (lowest Type II error) on the validation set. The model is designed to prioritize the maximization of the Recall score due to the belief that it is more crucial to minimize the Type II error rather than just increasing the Accuracy/F1 score of the model.

IV. RESULTS AND DISCUSSION

The findings of this study demonstrate that the proposed approach can effectively identify relations between medical entities without the need for supervision or manual intervention. The proposed method was applied to a collection of abstracts obtained from PubMed, and the results showed that the proposed unsupervised approach outperformed the previous supervised methods in terms of overall precision and F1-score on both the ChemProt and DDI datasets.

Theoretical results obtained from relation extraction can be used to develop more accurate and efficient algorithms that can automatically extract relations between biomedical entities from large-scale literature data. This has significant implications for drug discovery, disease diagnosis, and treatment, as it can facilitate the identification of new drug targets and drug-drug interactions, which can lead to the development of new treatments and cures for complex diseases. Additionally, the identification of disease biomarkers can aid in the early detection and diagnosis of diseases and provide insights into disease mechanisms, leading to more effective treatments and improved patient outcomes. Ultimately, the practical use of theoretical results obtained from relation extraction can have a profound impact on the field of biomedical research and accelerate the pace of scientific discovery, leading to new insights and discoveries that can improve human health and well-being. The pipeline can be used to extract relationships between biomedical entities in a variety of scenarios, facilitating the discovery of new relationships and insights in the biomedical domain.

A. Datasets

For training, the proposed model utilized a self-created dataset to train the BERT model. For testing and evaluation, it employed two widely recognized benchmark datasets in the biomedical domain: the DDI [29] dataset and the ChemProt [30] dataset, both of which involve multiclass classifications. Table IV presents the statistics of the DDI and ChemProt datasets.

B. Results on Benchmark Datasets

The performance of the proposed unsupervised approach for relation extraction from biomedical literature was evaluated on two well-known benchmark datasets, ChemProt and DDI. The results of the evaluation are presented in Table V, and the proposed approach was compared to a BERT model that underwent fine-tuning using supervised data manually annotated for the two datasets.

The results demonstrate that the proposed approach outperforms the most advanced model on the CHEMPROT dataset, achieving an F1-score of 85.8 compared to 75.14 for

the supervised approach. On the DDI dataset, the proposed approach achieves an F1-score of 88.5, compared to 70.2 for the supervised approach. Furthermore, the proposed unsupervised approach achieves a Recall score of 87.5 on the ChemProt dataset, compared to 75.09 for the supervised approach and a recall of 93.5 vs. 73.4 on DDI dataset, indicating that the proposed approach is more effective at correctly identifying relevant relationships. Fig. 8 provides a comparison of the results on the two datasets, highlighting the competitive performance of the proposed unsupervised approach.

TABLE IV. THE STATISTICS OF THE DDI AND THE CHEMPROT DATASETS

Dataset	DDI	ChemProt
Abstract	191	800
Positive relations	979	3458
Negative relations	4737	10,540

TABLE V. COMPARISON WITH EXISITNG MODELS ON CHEMPROT AND DDI DATASETS

Dataset	*Method*	*Precision*	*Recall*	*F1-Score*
ChemProt	Proposed method	83.15	87.5	85.8
	[31]	75.20	75.09	75.14
DDI	Proposed method	84.2	93.5	88.5
	[32]	-	73.4	70.2

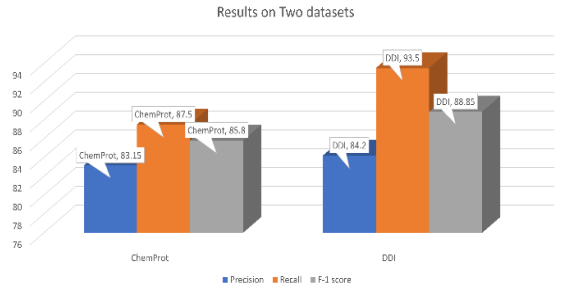

Fig. 8. Comparison of the results on the two benchmark datasets.

The results demonstrate that the proposed unsupervised approach for relation extraction from biomedical literature can strongly compete with the most advanced fully supervised systems, providing a scalable and efficient solution for relation extraction from biomedical literature. The success of the proposed approach in achieving performance equivalent to a fully supervised model highlights its potential to facilitate the discovery of new relationships and insights in the biomedical domain.

C. Results on Self-created Dataset

By analyzing Fig. 9, it can be noted that the model succeeded in its main goal to extract specific targeted biomedical relations from sentences and that the most frequent relations were cause and treat. The number of tagged genes, chemicals, proteins, and diseases in the abstracts were 27123, 30289, 25013 and 19471 respectively, which indicates the complexity and richness of the biomedical text, which can be challenging to process and extract information from without automated methods.

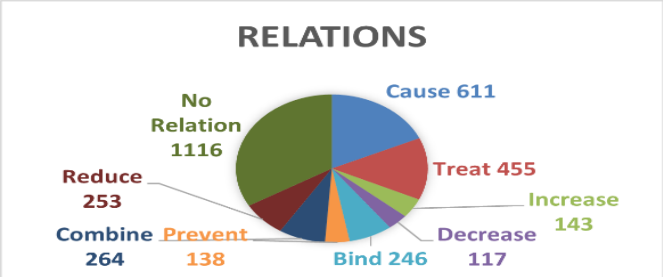

Fig. 9. Extracted relations count from self-created dataset.

The number of relations, as shown in Fig. 9, extracted by the model, mapped to 'cause' and 'treat' are significant, indicating the importance of these relationships in biomedical research. The numbers of relations extracted for increase, decrease, bind, combine and reduce demonstrate the model's ability to identify a variety of relationships beyond simple causation and treatment.

The number of sentences that have no relation is also an expected outcome, as not all sentences in the biomedical text contain explicit relationships between entities. However, it is still important to accurately identify these sentences to avoid false positives in downstream analyses. Overall, the successful extraction of targeted biomedical relations from the abstracts demonstrates the potential of automated methods for relation extraction in biomedical text mining. These methods can help researchers efficiently process and extract information from the vast amount of biomedical literature available, accelerating scientific discoveries and advancing healthcare.

V. CONCLUSION

The proposed pipeline based on the state-of-the-art BERT model offers a promising solution to the challenge of relation extraction from biomedical literature, providing a scalable and efficient approach that eliminates the need for human intervention or manual curation and which would reduce manpower and time consumption and automatically extract biomedical entities association data sets from large-scale literature data. The pipeline first retrieves articles (mainly abstracts) from PubMed according to specific queries and the extracted abstracts are then preprocessed to precisely extract the relation between medical entities. The preprocessed data are first presented in the form of subject-verb-object using part of speech tagger, with the resulting verbs and relationship types of interest embedded in a vector space. Each verb is mapped to the relation type that has the highest score after calculating the similarity score. The generated data set is used to fine-tune a BERT model to carry out relation extraction.

The significance of this study lies in its ability to reduce manpower and time consumption associated with relation extraction, making it more feasible and cost-effective. The use of the BERT model, which has demonstrated state-of-the-art performance in natural language processing tasks, provides a promising solution to the challenge of relation extraction from biomedical literature. The empirical comparison conducted in this study validates the effectiveness of the approach, demonstrating superior results compared to previous works that relied on supervised learning. Our unsupervised approach outperformed the supervised approach found in the literature in

the overall precision and F1-score on both the ChemProt and DDI datasets. In the ChemProt dataset, our method achieved a precision of 83.15 and an F1-score of 85.8, while the supervised method achieved only 75.2 and 75.14, respectively. Furthermore, our method achieved an F1-score of 88.5, surpassing the previous supervised methods scoring 70.2 in the DDI dataset.

The findings of this study have significant implications for future research and applications in this area, such as its potential scalability to larger datasets and its potential for automated extraction of biomedical relations from a wide range of scientific literature. This pipeline can be used to extract relationships between biomedical entities in a variety of scenarios, facilitating the discovery of new relationships and insights in the biomedical domain. By automating the extraction of biomedical relations from a wide range of scientific literature, researchers will be able to identify new relationships and insights that were previously hidden or difficult to discover. This will enable researchers to develop new hypotheses and accelerate the discovery of new treatments and cures for complex diseases and by automating the extraction of biomedical relations, we can accelerate the pace of scientific discovery and improve our understanding of complex diseases. The approach described in the paper uses Open Information Extraction (OIE) techniques to identify interaction words, this method can be applied to other domains and types of data that have a similar structure, such as news articles, social media posts, and legal documents, among others. The techniques used could potentially be adapted and applied to other domains and types of data with similar structures. However, the effectiveness of the approach may depend on the specific characteristics of the data being used, and further research is needed to fully understand its applicability and limitations in different contexts.

REFERENCES

[1] "PubMed", National Library of Medicine(US), National Center for Biotechnology Information, January 1996. [Online]. Available: http://pubmed.ncbi.nlm.nih.gov/. [Accessed March 2023].

[2] Mausam, M. Open Information Extraction Systems and Downstream Applications. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, NY, USA, 15 July 2016.

[3] Otter DW, Medina JR , Kalita JK. A survey of the usages of deep learning for natural language processing. IEEE Trans Neural Netw Learn Syst. 2020.

[4] Mausam, M. Open Information Extraction Systems and Downstream Applications. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, NY, USA, 15 July 2016.

[5] Kenton, J. D. M.-W. C., & Toutanova, L. K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Paper presented at the Proceedings of NAACL-HLT Minneapolis, USA, June 3, 2019.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie, Subbiah, Jared Kaplan, Prafulla Dhariwal, ArvindNeelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. ArXiv, abs/2005.14165.

[7] Diego Marcheggiani, I. T. (2016). Discrete-state variational autoencoders for joint discovery and factorization of relations. Transactions of the Association for Computational Linguistics, 231–244.

[8] Zhu, Y.; Li, L.; Lu, H.; Zhou, A.; Qin, X. Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions. J. Biomed. Inform. 2020, 106, 103451.

[9] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. arXiv preprint arXiv:1706.04115

[10] You Z H, Huang Z A, Zhu Z, et al. PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction. PLoS computational biology, 2017, 13(3): e1005455.

[11] Zaikis, D., Vlahavas, I. (2021). TP-DDI: Transformer-based Pipeline for the Extraction of Drug-Drug Interactions. Artif. Intell. Med., 119, 102153.

[12] Gu J, Sun F, Qian L, Zhou G. Chemical-induced disease relation extraction via convolutional neural network. Database (Oxford). 2017;2017:bax024.

[13] Mondal I. BERTChem-DDI: Improved Drug-Drug Interaction Prediction from text using Chemical Structure Information. arXiv preprint arXiv:2012.11599, 2020.

[14] Zhu Y, Li L, Lu H, et al. Extracting Drug-Drug Interactions from Texts with BioBERT and Multiple Entity-aware Attentions. Journal of Biomedical Informatics, 2020: 103451.

[15] Y. Peng, C.H. Wei, Z. Lu Improving chemcial disease relation extraction with rich features and weakly labeled dataJ. Cheminform., 8 (2016), p. 53.

[16] Chen X, Yan C C, Zhang X, et al. WBSMDA: within and between score for MiRNA-disease association prediction. Scientific reports, 2016, 6: 21106.

[17] Sänger M, Leser U. Large-scale entity representation learning for biomedical relationship extraction. Bioinformatics, 2020.

[18] Y. Zhang, H. Lin, Z. Yang, J. Wang, S. Zhang, Yuanyuan, L. Yang A hybrid model based on neural networks for biomedical relation extraction J. Biomed. Inform., 81 (2018), p. 83.

[19] Dina A. Salem, Breast Cancer Patients Using Mobile Applications: An Automated Biomedical Literature Curation Model (BLCM) | IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/10145269.

[20] Neumann, M. K. (2019). ScispaCy: fast and robust models for biomedical natural language processing. Proceedings of the18th BioNLP Workshop and Shared Task, 319-327.

[21] Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc."

[22] M. Peters, M. N. (2018). Deep contextualized word representations. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2227–2237.

[23] Howard Jeremy and Ruder Sebastian. 2018. Universal language model fine-tuning for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 328–339.

[24] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," arXiv preprint arXiv:1609.08144, 2016.

[25] Yang, Y.; Uy, M.C.S.; Huang, A. FinBERT: A pretrained language model for financial communications. arXiv 2020, arXiv:2006.08097. [Google Scholar].

[26] Dumitrescu, S.D.; Avram, A.M.; Pyysalo, S. The birth of Romanian BERT. arXiv 2020, arXiv:2009.08712.

[27] Jahan, M.S.; Beddiar, D.R.; Oussalah, M.; Arhab, N. Hate and Offensive language detection using BERT for English Subtask A. In Proceedings of the FIRE 2021: Forum for Information Retrieval Evaluation, Gandhinagar, India, 13–17 December 2021.

[28] Keya, A.J.; Wadud, M.A.H.; Mridha, M.F.; Alatiyyah, M.; Hamid, M.A. AugFake-BERT: Handling Imbalance through Augmentation of Fake

News Using BERT to Enhance the Performance of Fake News Classification. Appl. Sci. 2022, 12, 8398.

[29] M. Herrero-Zazo, I. S.-B. (2013). The DDI corpus: an annotated corpus with pharmacological substances and drug-drug interactions. Journal of Biomedical Informatics, 914-920.

[30] Kringelum J, K. S. (2016). ChemProt-3.0: a global chemical biology diseases mapping. Database, 1-7.

[31] Lee J, Y. W. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 1234-1240.

[32] Luo L, Y. Z. (2020). A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature. Biomed Inform 2020.