RESEARCH

Open Access

GPDminer: a tool for extracting named entities and analyzing relations in biological literature

Yeon-Ji Park[1], Geun-Je Yang[1], Chae-Bong Sohn[1]* and Soo Jun Park[2]*

*Correspondence:
cbsohn@kw.ac.kr; psj@etri.re.kr

[1] Department of Electronics and Communications Engineering, Kwangwoon University, 20 Gwangun-ro, Seoul 01897, Republic of Korea
[2] Welfare & Medical ICT Research Department, Electronics and Telecommunications Research Institute, 218 Gajeong-ro, Daejeon 34129, Republic of Korea

Abstract

Purpose: The expansion of research across various disciplines has led to a substantial increase in published papers and journals, highlighting the necessity for reliable text mining platforms for database construction and knowledge acquisition. This abstract introduces GPDMiner(Gene, Protein, and Disease Miner), a platform designed for the biomedical domain, addressing the challenges posed by the growing volume of academic papers.

Methods: GPDMiner is a text mining platform that utilizes advanced information retrieval techniques. It operates by searching PubMed for specific queries, extracting and analyzing information relevant to the biomedical field. This system is designed to discern and illustrate relationships between biomedical entities obtained from automated information extraction.

Results: The implementation of GPDMiner demonstrates its efficacy in navigating the extensive corpus of biomedical literature. It efficiently retrieves, extracts, and analyzes information, highlighting significant connections between genes, proteins, and diseases. The platform also allows users to save their analytical outcomes in various formats, including Excel and images.

Conclusion: GPDMiner offers a notable additional functionality among the array of text mining tools available for the biomedical field. This tool presents an effective solution for researchers to navigate and extract relevant information from the vast unstructured texts found in biomedical literature, thereby providing distinctive capabilities that set it apart from existing methodologies. Its application is expected to greatly benefit researchers in this domain, enhancing their capacity for knowledge discovery and data management.

Keywords: Natural language process, Database curation, Text mining, Mining tool, Biomedical domain, Named-entity recognition, Relation extraction

Introduction

The rapid advancement of contemporary research and academia has ushered in an era characterized by the astonishing production rate of new papers and academic resources [1]. While this phenomenon accelerates the dissemination of knowledge, it concurrently amplifies the challenges researchers face in maintaining and updating specialized knowledge in specific domains [2]. This issue is particularly pronounced in the field of biomedicine, attributed to the complexity of the domain and the intricate interconnectivity among various entities [3]. Consequently, researchers require considerable effort to identify and analyze pertinent information in this dynamic research environment [4]. In this context, the significance of text mining, aimed at extracting and analyzing information from vast amounts of unstructured text data, is on the ascent [5]. There is an acute demand for robust analysis in the realm of text mining, coupled with the development of platforms for database construction and knowledge acquisition [6]. While information retrieval approaches have proven efficacious for particular subjects, the continual surge in publications raises concerns regarding the efficiency of conventional search and knowledge acquisition methodologies. In biomedical research, considerable attention has been devoted to extracting information related to specific entities [7]. However, a wealth of invaluable information remains untapped within unstructured texts [8].

Most of the text mining tools available to date either focus predominantly on specific topics or fall short in processing data with intricate structures [9]. Such limitations restrict the practical applications of text mining. To address these shortcomings, in this paper, we introduce a novel text mining platform: Gene, Protein, and Disease Miner (GPDMiner). This platform is designed to automatically search, extract, and analyze the voluminous text data emerging in the field of biomedicine, enabling users to rapidly and efficiently acquire the information they seek. At the core of GPDMiner is its capability to leverage the extensive repository of papers available via PubMed. Specifically, this platform emphasizes querying specific topics from large databases and elucidating the relationships between biomedical entities in search results, thereby effectively visualizing interconnected information. Additionally, GPDMiner enhances the overall user experience by offering the ability to save and analyze results in various formats, including Excel and images.

It's crucial to recognize that Named Entity Recognition (NER) forms an integral component of this platform. NER involves identifying specific entities, such as people, date and time information, and biological proteins, within textual data. Since the term NER was first introduced in MUC-6, there has been a growing interest in natural language processing across various domains, leading to numerous studies related to NER [10–16]. In addition, different researchers have advocated for definitions or restrictions on named entities (NE) in various ways [17, 18]. Currently, NE is classified into two categories: general NE and domain-specific NE, with our focus on the latter. NER tasks have traditionally involved rule-based approaches, unsupervised learning approaches, and feature-based supervised learning approaches [18]. However, with the rapid advancement of deep learning, new possibilities have emerged for executing NER tasks using deep learning-based approaches that effectively learn intricate and sophisticated features [19]. By integrating these cutting-edge deep learning technologies into GPDMiner, we have enabled the extraction and recognition of domain-specific entities from biomedical texts.

Such research endeavors aim to bridge the gap between information overload and knowledge discovery in the realm of biomedical research. By providing a user-friendly tool to both researchers and non-experts, GPDMiner seeks to assist individuals in effectively harnessing the wealth of information hidden within biomedical texts. This paper aims to contribute to the rapid expansion and progress of knowledge acquisition and analysis in the burgeoning field of biomedicine by examining the potential and significance of GPDMiner as a text mining platform in biomedical research.

GPDMiner platform description

The exponential increase in research and publications across various academic disciplines underscores the importance of sophisticated and reliable platforms for text mining and knowledge acquisition [20]. In the biomedical field specifically, the complexity of numerous articles and journals presents considerable challenges for information retrieval and knowledge extraction tasks [21]. As an attempt to address these challenges, we have developed a new platform called GPDMiner by integrating novel technologies into existing research [22, 23]. By integrating statistical and dictionary-based analysis systems with Google's Bidirectional Encoder Representations from Transformers (BERT) technology, GPDMiner offers the capability to analyze the interrelationships among genes, proteins, and diseases. The core components of GPDMiner include:

1. **Entity recognition and relationship extraction:** Utilizing deep learning models for recognizing complex biomedical entities and discerning their relationships.
2. **Data source integration:** Provides query and search functionalities from various sources, including PubMed and the US Patent database.
3. **Relationship analysis:** Extracts relationships based on the influence index of reference literature, visualizing both the union and intersection of the results.
4. **Comprehensive and reliable results:** Combining traditional statistical parsers with BERT-based learning approaches for enhanced reliability.

Figure 1 provides a concise overview of the GPDMiner pipeline. Subsequent sections of this paper offer detailed descriptions of GPDMiner's individual components and architecture, encompassing how the system collects papers, analyzes data, visualizes results, and stores them. In conclusion, GPDMiner is an innovative platform that utilizes advanced text mining techniques, allowing researchers to effectively analyze complex entities and relationships in the biomedical domain. Such capabilities present a crucial solution to address the rapidly expanding complexity and volume of data in the contemporary biomedical research environment.

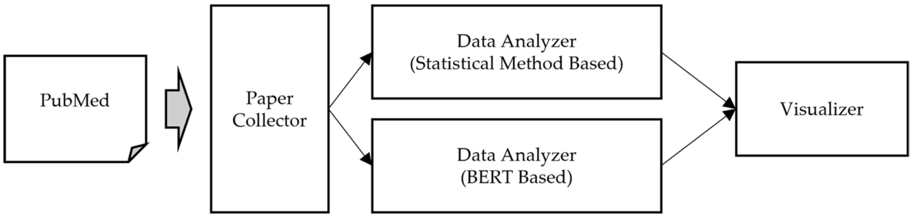

PubMed
Paper Collector
Data Analyzer (Statistical Method Based)
Data Analyzer (BERT Based)
Visualizer

Fig. 1 Pipeline of GPDMiner

GPDMiner system architecture

In this section, we introduce the architecture of GPDMiner, a novel system for entity and relationship recognition. GPDMiner is bifurcated into a Java-based main client program and a server program equipped with a Python-based BERT artificial intelligence engine. The system utilizes the PostgreSQL database for its construction. With its capability to deliver intricate and rapid analysis in the field of biomedical sciences, GPDMiner aims to alleviate researchers' efforts, saving time and enhancing accuracy.

The system architecture of GPDMiner is structured as follows. Figure 2 depicts the flowchart representing module-wise operations during mining execution between the GPDMiner client and the BERT server. Due to differences in programming languages, the design relies on Socket communication, transmitting data in a JSON format. This ensures seamless interaction between the client and server, enabling users to efficiently conduct PubMed searches and analyze the desired literature information.

The client searches for literature on PubMed, selects relevant publications, and transmits these data to the server using both the 'Reserve Mining' and 'Start Mining' functionalities. The server processes the data provided by the client by utilizing pretrained weights within the BERT server to conduct NER and Relation Extraction (RE). The server then transmits the analytical results to the client, which subsequently stores them in the database. These stored outcomes are visualized through "Mined Information," granting users the convenience to review analytical results, thereby informing their research and decision-making processes.

Figure 3 illustrates the detailed workflow of each module in GPDMiner. This flowchart elucidates the process by which the BERT engine, integrated as a server to the preexisting client, extracts and visualizes the relationships among diseases, genes/proteins, and gene/protein-diseases for user consumption.

Within the GPDMiner client, information from selected publications is utilized to obtain NE and RE results based on both statistical methods and BERT-based approaches. These results are then stored in a database (DB) table. To visualize the stored analytical outcomes, the system generates a data list by fetching analysis data from the DB using SQL queries, based on the desired PubMed ID (PMID). Subsequent to this, data tables dedicated for visualization are produced. During this phase, any duplicate entries in these visualization tables undergo a refinement process to ensure uniqueness. Afterwards, the union and intersection data sets of the statistical method-based analysis and

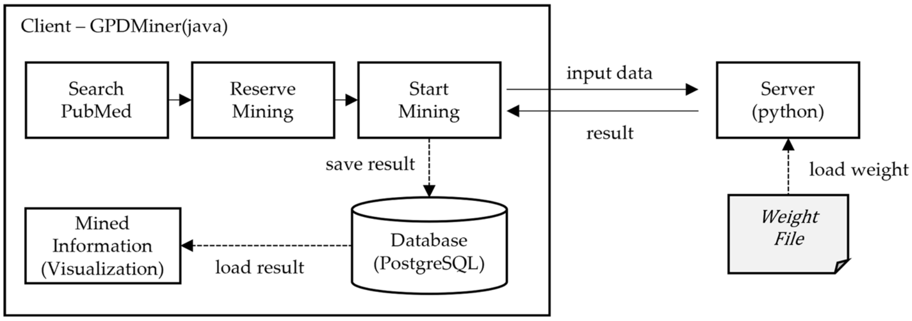


Fig. 2 Overview of GPDMiner

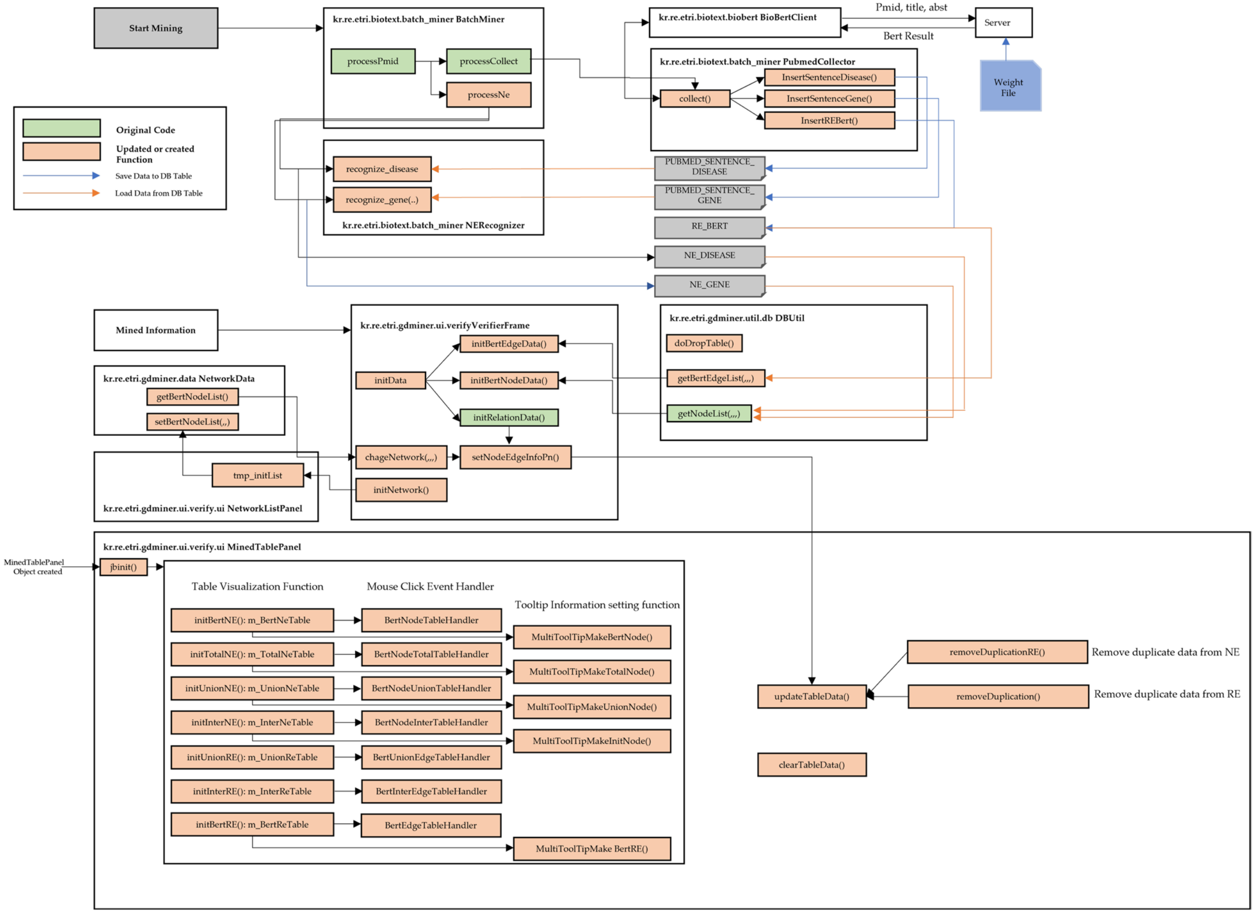

Start Mining
kr.re.etri.biotext.batch_miner BatchMiner
processPmid
processCollect
processNe
kr.re.etri.biotext.biobert BioBertClient
Pmid, title, abst
Bert Result
Server
Weight File
kr.re.etri.biotext.batch_miner PubmedCollector
collect()
InsertSentenceDisease()
InsertSentenceGene()
InsertREBert()
Original Code
Updated or created Function
Save Data to DB Table
Load Data from DB Table
recognize_disease
recognize_gene(..)
kr.re.etri.biotext.batch_miner NERecognizer
PUBMED_SENTENCE_ DISEASE
PUBMED_SENTENCE_ GENE
RE_BERT
NE_DISEASE
NE_GENE
Mined Information
kr.re.etri.gdminer.ui.verify VerifierFrame
initData
initBertEdgeData()
initBertNodeData()
initRelationData()
chageNetwork(,,,)
setNodeEdgeInfoPn()
initNetwork()
kr.re.etri.gdminer.util.db DBUtil
doDropTable()
getBertEdgeList(,,,)
getNodeList(,,,)
kr.re.etri.gdminer.data NetworkData
getBertNodeList()
setBertNodeList(,,)
tmp_initList
kr.re.etri.gdminer.ui.verify.ui NetworkListPanel
kr.re.etri.gdminer.ui.verify.ui MinedTablePanel
MinedTablePanel Object created
jbinit()
Table Visualization Function
Mouse Click Event Handler
Tooltip Information setting function
initBertNE(): m_BertNeTable
initTotalNE(): m_TotalNeTable
initUnionNE(): m_UnionNeTable
initInterNE(): m_InterNeTable
initUnionRE(): m_UnionReTable
initInterRE(): m_InterReTable
initBertRE(): m_BertReTable
BertNodeTableHandler
BertNodeTotalTableHandler
BertNodeUnionTableHandler
BertNodeInterTableHandler
BertUnionEdgeTableHandler
BertInterEdgeTableHandler
BertEdgeTableHandler
MultiToolTipMakeBertNode()
MultiToolTipMakeTotalNode()
MultiToolTipMakeUnionNode()
MultiToolTipMakeInitNode()
MultiToolTipMake BertRE()
updateTableData()
clearTableData()
removeDuplicationRE()
removeDuplication()
Remove duplicate data from NE
Remove duplicate data from RE

Fig. 3 Detailed module-wise workflow of GPDMiner

BERT-based analysis are generated. Through these datasets, an intricate analytical procedure ensues, culminating in the final visualization. This resultant visual representation aids users in gaining a clearer understanding of the interrelations among diseases, genes/proteins, and gene/protein-diseases.

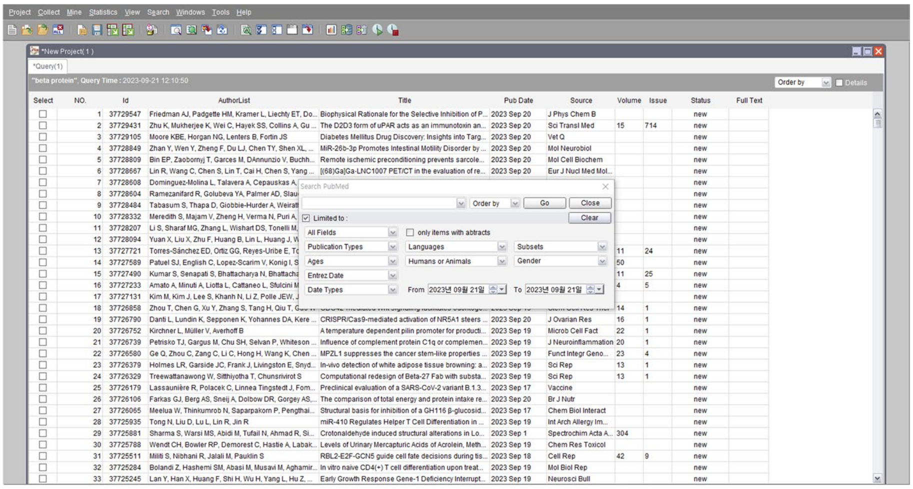
The architecture of GPDMiner offers a streamlined analytical environment to its users. The integration of the BERT engine amplifies its capabilities, allowing for more precise relation extraction and visualization [24]. Such features hold significant value in assisting researchers and decision-making processes within the biomedical field. It's anticipated that these functionalities will contribute to the complex data analysis and knowledge acquisition [25].

Paper collector

GPDMiner's search interface conforms to PubMed's standard search methodology, with the PubMed collector playing a pivotal role. This collector employs the Entrez Programming Utilities to search for and succinctly accumulate abstract information for specific queries.

Figure 4 depicts the search result interface of the GPDMiner Collector. Once a PubMed search is executed, the retrieved information is stored in a database. Subsequent analytical procedures extract data from this database, facilitating intricate processing tasks. Such a methodology proves crucial in efficiently (https://www.overleaf.com/project/65cf152f70616ff65e6b1f65) managing the desired abstract data, and in promptly furnishing necessary information during the analytical phase.

Users can view search results under the 'Projects' tab. Each project comprises multiple tabs, each representing a distinct query. Within the project tab, users can review, requery, and amend the search results. Tab names are auto-generated based on the query string, although they can be renamed per user preference. Each tab displays query information, such as the query string, query timestamp, and sorting order, with the resulting abstracts presented in a tabular format. To maintain up-to-date data, either the entire



"beta protein", Query Time : 2023-09-21 12:10:50

Select	NO.	Id	AuthorList	Title	Pub Date	Source	Volume	Issue	Status	Full Text
□	1	37729547	Friedman AJ, Padgette HM, Kramer L, Liechty ET, Do...	Biophysical Rationale for the Selective Inhibition of P...	2023 Sep 20	J Phys Chem B			new	
□	2	37729431	Zhu K, Mukherjee K, Wei C, Hayek SS, Collins A, Gu ...	The D2D3 form of uPAR acts as an immunotoxin an...	2023 Sep 20	Sci Transl Med	15	714	new	
□	3	37729105	Moore KBE, Horgan NG, Lenters B, Fortin JS	Diabetes Mellitus Drug Discovery: Insights into Targ...	2023 Sep 20	Vet Q			new	
□	4	37728849	Zhan Y, Wen Y, Zheng F, Du LJ, Chen TY, Shen XL, ...	MiR-26b-3p Promotes Intestinal Motility Disorder by ...	2023 Sep 20	Mol Neurobiol			new	
□	5	37728809	Bin EP, Zaobornyj T, Garces M, DAnnunzio V, Buchh...	Remote ischemic preconditioning prevents sarcole...	2023 Sep 20	Mol Cell Biochem			new	
□	6	37728667	Lin R, Wang C, Chen S, Lin T, Cai H, Chen S, Yang ...	[(68)Ga]Ga-LNC1007 PET/CT in the evaluation of re...	2023 Sep 20	Eur J Nucl Med Mol...			new	
□	7	37728608	Dominguez-Molina L, Talavera A, Cepauskas A,						new	
□	8	37728604	Ramezanifard R, Golubeva YA, Palmer AD, Slau						new	
□	9	37728484	Tabasum S, Thapa D, Giobbie-Hurder A, Weirat						new	
□	10	37728332	Meredith S, Majam V, Zheng H, Verma N, Puri A,						new	
□	11	37728207	Li S, Sharaf MG, Zhang L, Wishart DS, Tonelli M,						new	
□	12	37728094	Yuan X, Liu X, Zhu F, Huang B, Lin L, Huang J, W						new	
□	13	37727721	Torres-Sánchez ED, Ortiz GG, Reyes-Uribe E, To				11	24	new	
□	14	37727589	Patuel SJ, English C, Lopez-Scarim V, Konig I, S				50		new	
□	15	37727490	Kumar S, Senapati S, Bhattacharya N, Bhattacha				11	25	new	
□	16	37727233	Amato A, Minuti A, Liotta L, Cattaneo L, Sfulcini M				4	5	new	
□	17	37727131	Kim M, Kim J, Lee S, Khanh N, Li Z, Polle JEW, J						new	
□	18	37726858	Zhou T, Chen G, Xu Y, Zhang S, Tang H, Qiu T, Gao ...	[illegible]	[illegible]	[illegible]	14	1	new	
□	19	37726790	Danti L, Lundin K, Sepponen K, Yohannes DA, Kere ...	CRISPR/Cas9-mediated activation of NR5A1 steers ...	2023 Sep 20	J Ovarian Res	16	1	new	
□	20	37726752	Kirchner L, Müller V, Averhoff B	A temperature dependent pilin promoter for producti...	2023 Sep 19	Microb Cell Fact	22	1	new	
□	21	37726739	Petrisko TJ, Gargus M, Chu SH, Selvan P, Whiteson ...	Influence of complement protein C1q or complemen...	2023 Sep 19	J Neuroinflammation	20	1	new	
□	22	37726580	Ge Q, Zhou C, Zang C, Li C, Hong H, Wang K, Chen ...	MPZL1 suppresses the cancer stem-like properties ...	2023 Sep 19	Funct Integr Geno...	23	4	new	
□	23	37726379	Holmes LR, Garside JC, Frank J, Livingston E, Snyd...	In-vivo detection of white adipose tissue browning: a...	2023 Sep 19	Sci Rep	13	1	new	
□	24	37726329	Treewattanawong W, Sitthiyotha T, Chunsrivirot S	Computational redesign of Beta-27 Fab with substa...	2023 Sep 19	Sci Rep	13	1	new	
□	25	37726179	Lassaunière R, Polacek C, Linnea Tingstedt J, Fom...	Preclinical evaluation of a SARS-CoV-2 variant B.1.3...	2023 Sep 17	Vaccine			new	
□	26	37726106	Farkas GJ, Berg AS, Sneij A, Dolbow DR, Gorgey AS,...	The comparison of total energy and protein intake re...	2023 Sep 20	Br J Nutr			new	
□	27	37726065	Meelua W, Thinkumrob N, Saparpakorn P, Pengthai...	Structural basis for inhibition of a GH116 β-glucosid...	2023 Sep 17	Chem Biol Interact			new	
□	28	37725935	Tong N, Liu D, Lu L, Lin R, Jin R	miR-410 Regulates Helper T Cell Differentiation in ...	2023 Sep 19	Int Arch Allergy Im...			new	
□	29	37725881	Sharma S, Warsi MS, Abidi M, Tufail N, Ahmad R, Si...	Crotonaldehyde induced structural alterations in Lo...	2023 Sep 1	Spectrochim Acta A...	304		new	
□	30	37725788	Wendt CH, Bowler RP, Demorest C, Hastie A, Labak...	Levels of Urinary Mercapturic Acids of Acrolein, Meth...	2023 Sep 19	Chem Res Toxicol			new	
□	31	37725511	Militi S, Nibhani R, Jalali M, Pauklin S	RBL2-E2F-GCN5 guide cell fate decisions during tis...	2023 Sep 18	Cell Rep	42	9	new	
□	32	37725284	Bolandi Z, Hashemi SM, Abasi M, Musavi M, Aghamir...	In vitro naive CD4(+) T cell differentiation upon treat...	2023 Sep 19	Mol Biol Rep			new	
□	33	37725245	Lan Y, Han X, Huang F, Shi H, Wu H, Yang L, Hu Z, ...	Early Growth Response Gene-1 Deficiency Interrupt...	2023 Sep 19	Neurosci Bull			new	

☑ Limited to :

All Fields | □ only items with abtracts

Publication Types | Languages | Subsets

Ages | Humans or Animals | Gender

Entrez Date

Date Types | From 2023년 09월 21일 | To 2023년 09월 21일

Fig. 4 GPDMiner collector

list or selected items can be updated via a refresh command, with new entries automatically incorporated into the table.

This PubMed collector operates as a crucial component within the GPDMiner system, streamlining literature searches and data accumulation processes. This aids the efficiency of research and decision-making in the biomedical realm. Additionally, the flexible architecture of the GPDMiner system optimizes intricate query processing and result presentation, ensuring precise, swift access to information. The collected search and organized information are fed into GPDMiner's Analyzer. The following sections will describe how this data is used as input for NE and RE analysis.

Data analyzer

This section explains the functionality of GPDMiner's Analyzer, which processes the collected data. We present results from a standard statistical-based mining approach as well as a BERT-based mining approach.

Statistical method based mining

The statistical mining approach undergoes the following process: Upon inputting the selected abstracts, entities and their relations are manifested through a relational analyzer. Natural language processing techniques, including Part-of-Speech (POS) tagging and syntactic parsing, are employed to facilitate this. We apply a statistical NER method underpinned by the Maximum Entropy model.

Statistical NER is a technique designed to discern named entities within a text. Among statistical techniques, we utilize the Maximum Entropy model. This model predicts which word or phrase in a given text corresponds to a particular entity type [26]. Based on principles of information theory, the model is trained to a maximum entropy state by selecting the prediction with the most uncertainty from the available data. By considering various features (e.g., words, contextual information), the model calculates the likelihood of each word or phrase belonging to a particular entity type, selecting the entity type with the highest probability. Due to its consideration of diverse features, the Maximum Entropy model can yield nuanced and flexible outcomes.

Furthermore, syntactic parsing is employed for extracting relationships between named entities. Syntactic parsing decomposes text into the structure and components of sentences, analyzing grammatical relationships between words. Through parsing, we discern roles and relations of named entities within a sentence, generating a parse tree that hierarchically represents grammatical relationships between words. Analyzing labels of each node in this tree, entities are identified and corresponding words are extracted. Relations between named entities are then deduced and extracted based on the syntactic parsing tree. By examining relationships inferred from the tree, we can comprehend the interactions between entities within the sentence.

BERT based mining

To attain more advanced outcomes, we present analytical findings of statistical-based mining techniques in conjunction with BERT-based mining methods.

BERT is a deep learning model grounded on the Transformer architecture, which has yielded breakthroughs in NLP. Leveraging the multi-head attention mechanism of the

Transformer, BERT simultaneously accounts for the interactions of all tokens in a sentence, introducing parallel processing capabilities not seen in traditional RNN-based models. This facilitates the analysis of intricate dependencies between tokens within a sentence. The bidirectional encoding feature of BERT can concurrently capture contexts on both sides of a specific word, playing a pivotal role in entity recognition and relationship extraction. In this study, we utilize BERT's capability to perform NE and RE in biomedical text. BERT was pre-trained with English Wikipedia and bookscorpus, but since the biomedical domain contains a variety of proper nouns and terms, BERT models aimed at understanding general-purpose languages do not perform well in biomedical test mining problems. To overcome these shortcomings, BioBERT is pre-trained using not only general texts but also medical literature collected from PubMed and PMC, which are specific corpora in the biomedical domain. We used this pre-trained model, BioBERT, to perform fine-tuning training on NE and RE tasks.

NE pertains to the process of identifying and classifying specific entities from text. In the realm of biomedical texts, this often revolves around entities such as genes, proteins, and diseases. Utilizing the BERT-based model, texts are tokenized, converted into numerical vectors, and contextually interpreted to discern the category of a given entity [27]. This procedure is integral to exacting precise information extraction from intricate biomedical contexts. RE delves into analyzing the semantic relationships between recognized entities. Bidirectional LSTMs and attention mechanisms are employed to excavate relationships between entity pairs within sentences, shedding light on complex biomedical relations. For instance, interactions between specific proteins and diseases or relations between genes and their traits are discerned. These NE and RE methodologies serve as core processes for information extraction and analysis from biomedical texts. Notably, by fine-tuning the pre-trained BERT model with a corpus tailored to the biomedical domain, we could more effectively capture domain-specific information.

For entity recognition, Fine-tuning was performed using the NCBI-Disease dataset and BC2GM dataset. Additionally, for relation recognition, the GAD dataset was employed as a fine-tuning dataset and integrated into the final relation recognition process. The detailed annotation counts for each dataset are presented in Table 1. Employing both statistical and BERT-based mining techniques, the analyzer of GPDMiner extracts vital information and discerns relationships from collected data.

Mining method functionalities

In this section, we elucidate the functionalities showcasing the mining results obtained through the aforementioned techniques. To illustrate each functionality, we display the analysis results of a query for 'beta protein' in a document with PMID 37563361.

Table 1 Statistics of datasets

Dataset	Entity type	Annotations
NCBI disease [28]	Disease	6881
BC2GM [29]	Gene/protein	13,270
GAD [30]	Gene-disease	5330

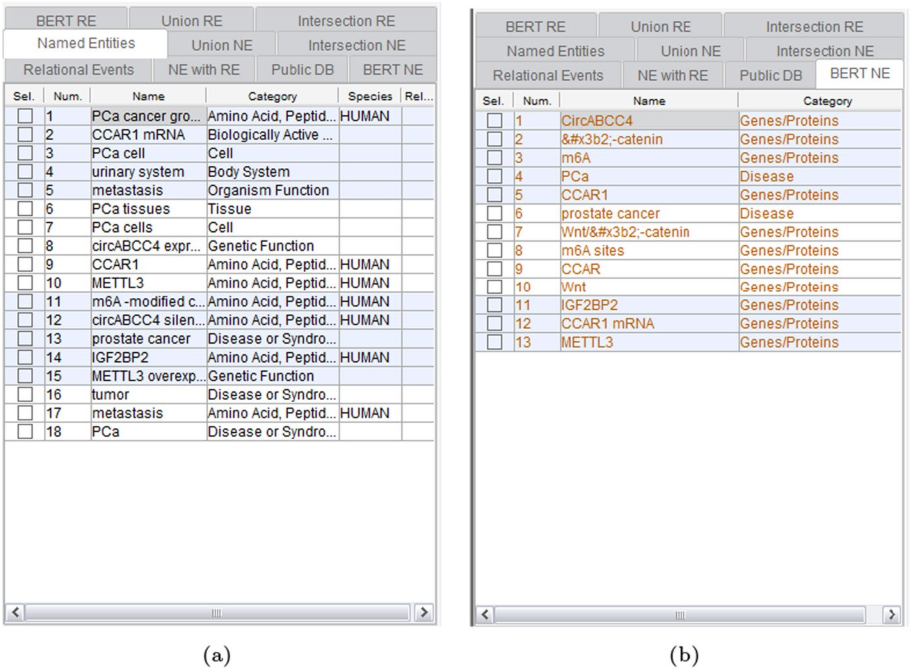
Firstly, Fig. 5 presents the results of NER by GPDMiner. Figure 5a illustrates the NER recognized solely by statistical and dictionary-based methods, which does not encompass those identified by BERT-based methods. Figure 5b highlights only the NER results identified through the BERT-based approach, distinctively marked in red to differentiate from the conventional results.

Subsequently, Fig. 6 illustrates the union and intersection results applied to the NER outcomes. Figure 6a features a functionality displaying both NE recognized through conventional statistical and dictionary-based methods, as well as the BERT-based approach. Figure 6b reveals NE common to both conventional and BERT-based methods. Different colors are employed to distinguish the results obtained from the conventional and BERT-based methods.

Moving forward, Fig. 7 displays the NE recognized through conventional statistical and dictionary-based methods, alongside the relevance-validated entries in public databases. This functionality provides users with relationships associated with the mined NER from public databases like HPRD and BioGRID.

Continuing, Fig. 8 presents the results of RE by GPDMiner. Figure 8a exhibits the relevance of entities recognized through conventional statistical and dictionary-based methods. This method displays directionality and allows for weighted display, taking into account the paper's impact factor. Figure 8b displays the gene/protein-disease relationship extracted through the BERT-based method. However, due to the engine's structure, the BERT-based relationship cannot show directionality and weights.

Lastly, Fig. 9a comprehensively depicts the relationships extracted through both conventional and BERT-based methods. Figure 9b focuses on relationships common to both methods. The conventional results are depicted in black, while the BERT-based



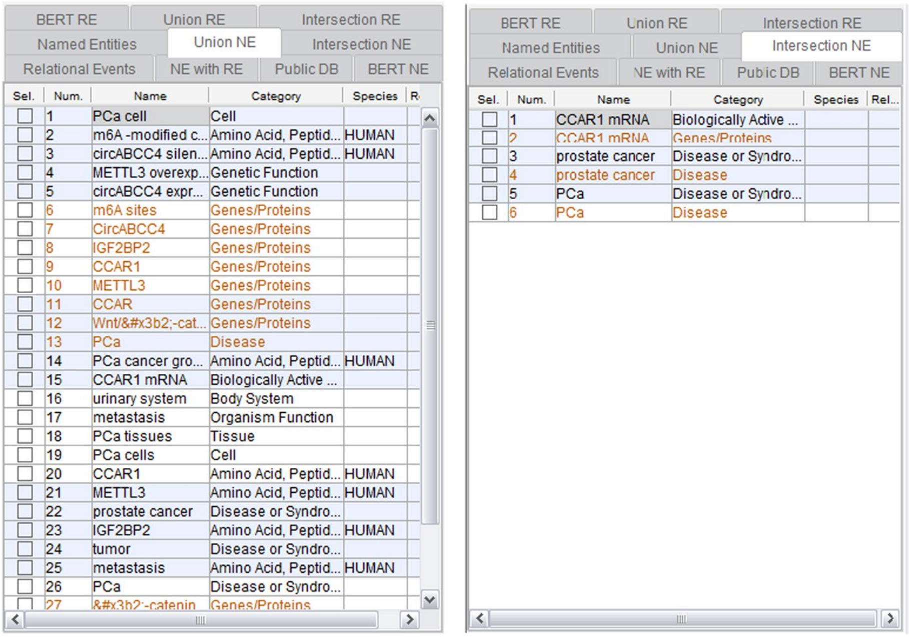
(a)

Sel.	Num.	Name	Category	Species	Rel...
□	1	PCa cancer gro...	Amino Acid, Peptid...	HUMAN	
□	2	CCAR1 mRNA	Biologically Active ...		
□	3	PCa cell	Cell		
□	4	urinary system	Body System		
□	5	metastasis	Organism Function		
□	6	PCa tissues	Tissue		
□	7	PCa cells	Cell		
□	8	circABCC4 expr...	Genetic Function		
□	9	CCAR1	Amino Acid, Peptid...	HUMAN	
□	10	METTL3	Amino Acid, Peptid...	HUMAN	
□	11	m6A -modified c...	Amino Acid, Peptid...	HUMAN	
□	12	circABCC4 silen...	Amino Acid, Peptid...	HUMAN	
□	13	prostate cancer	Disease or Syndro...		
□	14	IGF2BP2	Amino Acid, Peptid...	HUMAN	
□	15	METTL3 overexp...	Genetic Function		
□	16	tumor	Disease or Syndro...		
□	17	metastasis	Amino Acid, Peptid...	HUMAN	
□	18	PCa	Disease or Syndro...		

(b)

Sel.	Num.	Name	Category
□	1	CircABCC4	Genes/Proteins
□	2	&#x3b2;-catenin	Genes/Proteins
□	3	m6A	Genes/Proteins
□	4	PCa	Disease
□	5	CCAR1	Genes/Proteins
□	6	prostate cancer	Disease
□	7	Wnt/&#x3b2;-catenin	Genes/Proteins
□	8	m6A sites	Genes/Proteins
□	9	CCAR	Genes/Proteins
□	10	Wnt	Genes/Proteins
□	11	IGF2BP2	Genes/Proteins
□	12	CCAR1 mRNA	Genes/Proteins
□	13	METTL3	Genes/Proteins

Fig. 5 Entity recognition result of GPDMiner: **a** NER result using statistical and dictionary-based method; **b** NER result using BERT-based method



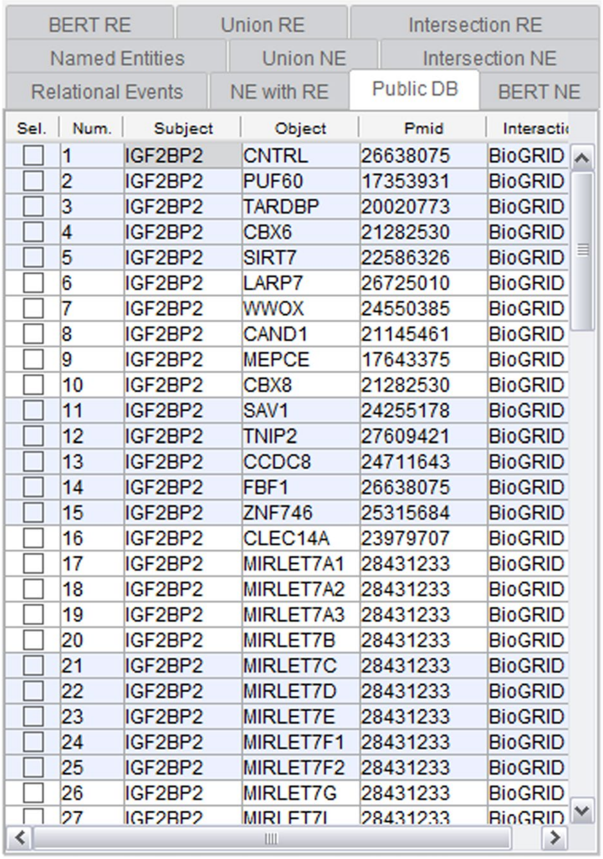
Sel.	Num.	Name	Category	Species	R
☐	1	PCa cell	Cell		
☐	2	m6A -modified c...	Amino Acid, Peptid...	HUMAN	
☐	3	circABCC4 silen...	Amino Acid, Peptid...	HUMAN	
☐	4	METTL3 overexp...	Genetic Function		
☐	5	circABCC4 expr...	Genetic Function		
☐	6	m6A sites	Genes/Proteins		
☐	7	CircABCC4	Genes/Proteins		
☐	8	IGF2BP2	Genes/Proteins		
☐	9	CCAR1	Genes/Proteins		
☐	10	METTL3	Genes/Proteins		
☐	11	CCAR	Genes/Proteins		
☐	12	Wnt/&#x3b2;-cat...	Genes/Proteins		
☐	13	PCa	Disease		
☐	14	PCa cancer gro...	Amino Acid, Peptid...	HUMAN	
☐	15	CCAR1 mRNA	Biologically Active ...		
☐	16	urinary system	Body System		
☐	17	metastasis	Organism Function		
☐	18	PCa tissues	Tissue		
☐	19	PCa cells	Cell		
☐	20	CCAR1	Amino Acid, Peptid...	HUMAN	
☐	21	METTL3	Amino Acid, Peptid...	HUMAN	
☐	22	prostate cancer	Disease or Syndro...		
☐	23	IGF2BP2	Amino Acid, Peptid...	HUMAN	
☐	24	tumor	Disease or Syndro...		
☐	25	metastasis	Amino Acid, Peptid...	HUMAN	
☐	26	PCa	Disease or Syndro...		
☐	27	&#x3b2;-catenin	Genes/Proteins		

(a)

Sel.	Num.	Name	Category	Species	Rel...
☐	1	CCAR1 mRNA	Biologically Active ...		
☐	2	CCAR1 mRNA	Genes/Proteins		
☐	3	prostate cancer	Disease or Syndro...		
☐	4	prostate cancer	Disease		
☐	5	PCa	Disease or Syndro...		
☐	6	PCa	Disease		

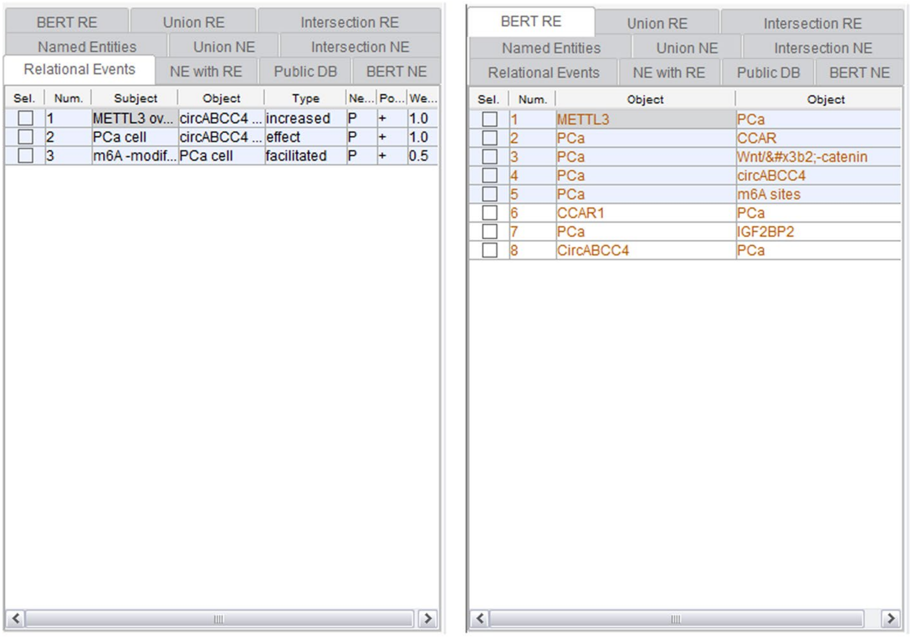
(b)

Fig. 6 Union and intersection results applied to NER outcomes: **a** Union applied to NER results; **b** Intersection applied to NER results



Sel.	Num.	Subject	Object	Pmid	Interacti
☐	1	IGF2BP2	CNTRL	26638075	BioGRID
☐	2	IGF2BP2	PUF60	17353931	BioGRID
☐	3	IGF2BP2	TARDBP	20020773	BioGRID
☐	4	IGF2BP2	CBX6	21282530	BioGRID
☐	5	IGF2BP2	SIRT7	22586326	BioGRID
☐	6	IGF2BP2	LARP7	26725010	BioGRID
☐	7	IGF2BP2	WWOX	24550385	BioGRID
☐	8	IGF2BP2	CAND1	21145461	BioGRID
☐	9	IGF2BP2	MEPCE	17643375	BioGRID
☐	10	IGF2BP2	CBX8	21282530	BioGRID
☐	11	IGF2BP2	SAV1	24255178	BioGRID
☐	12	IGF2BP2	TNIP2	27609421	BioGRID
☐	13	IGF2BP2	CCDC8	24711643	BioGRID
☐	14	IGF2BP2	FBF1	26638075	BioGRID
☐	15	IGF2BP2	ZNF746	25315684	BioGRID
☐	16	IGF2BP2	CLEC14A	23979707	BioGRID
☐	17	IGF2BP2	MIRLET7A1	28431233	BioGRID
☐	18	IGF2BP2	MIRLET7A2	28431233	BioGRID
☐	19	IGF2BP2	MIRLET7A3	28431233	BioGRID
☐	20	IGF2BP2	MIRLET7B	28431233	BioGRID
☐	21	IGF2BP2	MIRLET7C	28431233	BioGRID
☐	22	IGF2BP2	MIRLET7D	28431233	BioGRID
☐	23	IGF2BP2	MIRLET7E	28431233	BioGRID
☐	24	IGF2BP2	MIRLET7F1	28431233	BioGRID
☐	25	IGF2BP2	MIRLET7F2	28431233	BioGRID
☐	26	IGF2BP2	MIRLET7G	28431233	BioGRID
☐	27	IGF2BP2	MIRLET7I	28431233	BioGRID

Fig. 7 Public DB functionality



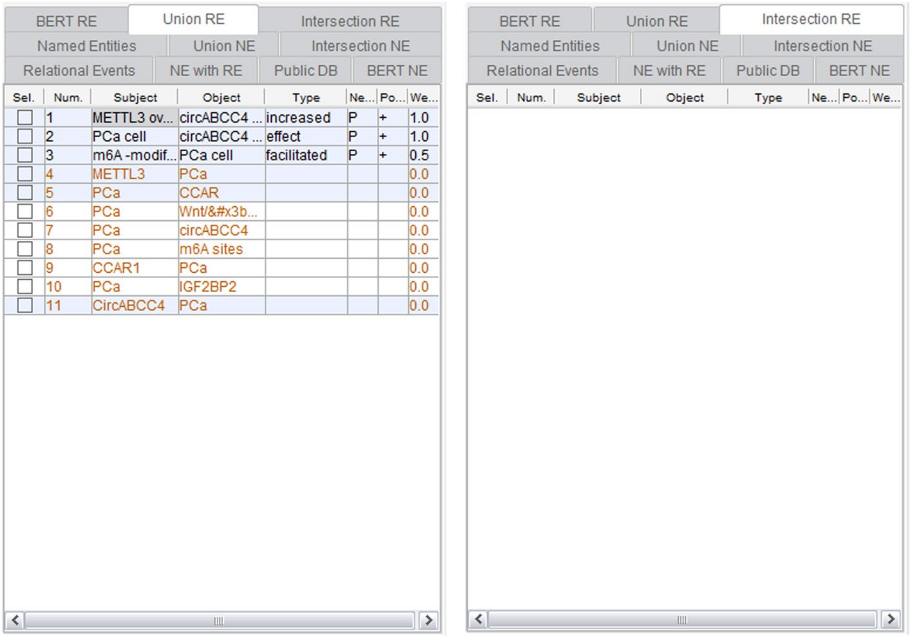
Sel.	Num.	Subject	Object	Type	Ne...	Po...	We...
☐	1	METTL3 ov...	circABCC4 ...	increased	P	+	1.0
☐	2	PCa cell	circABCC4 ...	effect	P	+	1.0
☐	3	m6A -modif...	PCa cell	facilitated	P	+	0.5

(a)

Sel.	Num.	Object	Object
☐	1	METTL3	PCa
☐	2	PCa	CCAR
☐	3	PCa	Wnt/&#x3b2;-catenin
☐	4	PCa	circABCC4
☐	5	PCa	m6A sites
☐	6	CCAR1	PCa
☐	7	PCa	IGF2BP2
☐	8	CircABCC4	PCa

(b)

Fig. 8 Relation extraction result of GPDMiner: **a** RE result using statistical and dictionary-based methods; **b** RE result using BERT-based method



Sel.	Num.	Subject	Object	Type	Ne...	Po...	We...
☐	1	METTL3 ov...	circABCC4 ...	increased	P	+	1.0
☐	2	PCa cell	circABCC4 ...	effect	P	+	1.0
☐	3	m6A -modif...	PCa cell	facilitated	P	+	0.5
☐	4	METTL3	PCa				0.0
☐	5	PCa	CCAR				0.0
☐	6	PCa	Wnt/&#x3b...				0.0
☐	7	PCa	circABCC4				0.0
☐	8	PCa	m6A sites				0.0
☐	9	CCAR1	PCa				0.0
☐	10	PCa	IGF2BP2				0.0
☐	11	CircABCC4	PCa				0.0

(a)

Sel.	Num.	Subject	Object	Type	Ne...	Po...	We...

(b)

Fig. 9 Union and intersection results applied to RE outcomes: **a** Union applied to RE results; **b** Intersection applied to RE results

outcomes are highlighted in red, facilitating clear differentiation between the two approaches. This comprehensive visualization enables users to proficiently analyze and comprehend the mining findings [31].

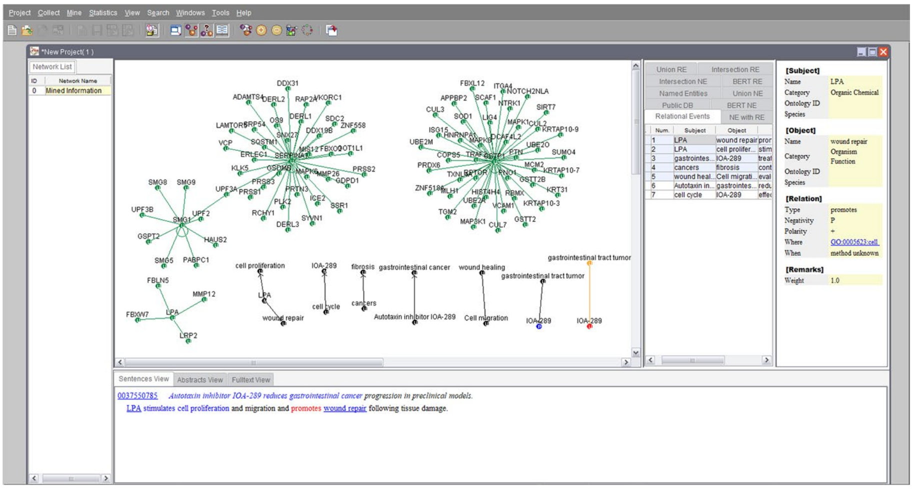
Visualizer

This section elucidates the process of visualizing analysis results using the data analyzer. Initiating with 'Mined Information', the mined results of the selected project are displayed.

Figure 10 illustrates the mining analysis results interface of GPDMiner. The client interface features a Network List table on the far left side, which presents a list of networks. Upon selecting a specific network, relevant information appears in the visualization screen, table screen, detailed information screen, and literature information screen on the right. The visualization screen presents data in a graphical format, while the table screen showcases NE and RE tables. By selecting specific cells in the NE and RE tables, corresponding details are presented in the detailed information screen and literature information screen. The detailed information screen presents users with a comprehensive view of an entity's name, category, ontology ID, type, and more. In the literature information section, the selected entity/relation information is highlighted in red, while other entity/relation information is marked in blue. Clicking on a PMID connects to the PubMed site, allowing access to the abstract associated with the information.

Mining results can be visualized in this way, and Fig. 10 illustrates the 'Mined Information' results screen for the literature with PMID 37550785, which is the result of a search query for "beta protein". To visualize the results of the BERT analysis, we have incorporated the following.

In addition to visualizing the results of relations based on Public DB and statistically derived methods, we have extended the visualization to include BERT-identified gene/protein and disease entities, as well as the relationships between gene/protein and disease. This enhancement aims to make high-quality information more readily comprehensible for users. Furthermore, to facilitate easy identification of information,



ID	Network Name
0	Mined Information

Num.	Subject	Object	
1	LPA	wound repair	pror
2	LPA	cell prolifer...	stim
3	gastrointes...	IOA-289	treat
4	cancers	fibrosis	cont
5	wound heal...	Cell migrati...	eval
6	Autotaxin in...	gastrointes...	redu
7	cell cycle	IOA-289	effec

[Subject]

Name	LPA
Category	Organic Chemical
Ontology ID	
Species	

[Object]

Name	wound repair
Category	Organism Function
Ontology ID	
Species	

[Relation]

Type	promotes
Negativity	P
Polarity	+
Where	GO:0005623:cell
When	method unknown

[Remarks]

Weight	1.0

0037550785 *Autotaxin inhibitor IOA-289 reduces gastrointestinal cancer progression in preclinical models.*
LPA stimulates cell proliferation and migration and promotes wound repair following tissue damage.

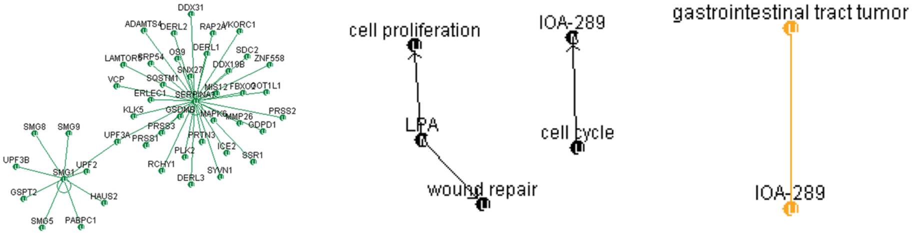
Fig. 10 The mining analysis result interface of GPDMiner

we have color-coded these relationship recognition outcomes in the graphical representation. Specifically, as depicted in Fig. 11, relationships based on Public DB are denoted in green, those derived from conventional statistical and dictionary-based methods in black, and BERT-based relationships in orange.

Through the 'Mined Information' process, users can effectively explore and interpret mining results using visualization tools and comprehensive data presentation. This user-centric interface significantly enhances the accessibility and utility of analysis outcomes, thereby supporting biomedical research and decision-making processes [32]. However, we acknowledge that more rigorous, systematic comparisons are needed to conclusively establish these benefits. Future research will focus on conducting such studies to provide empirical support for the enhanced accessibility and utility of GPDMiner.

Saver

GPDMiner offers a suite of intuitive functionalities tailored for the efficient storage and organization of complex analytical outputs. This tool establishes a structured framework that effectively streamlines and expedites data analysis and dissemination among researchers. The 'Save Mining Report to' feature empowers users with the capability to specify the report name, designate file pathways, and select an appropriate output format - either HTML or Excel. When opting for the Excel format, the system autonomously generates the file, encompassing three pivotal sections: 'Report', 'Relational Events', and 'Named Entities'. Notably, the management subsystem integrated within the 'Relational Events' and 'Named Entities' segments ensures an automated transition to a subsequent sheet upon reaching a threshold of 30,000 data entries per sheet. This innovative storage mechanism not only enhances efficiency but also provides the necessary scalability for handling extensive datasets, thereby reinforcing the platform's unwavering commitment to promoting research reproducibility and transparency. While primarily designed for researchers, this feature's utility extends to a broader demographic, offering advanced assistance for knowledge procurement and data mining endeavors, especially within the dynamic field of biomedical research where the need for efficient data management is paramount.



(a) (b) (c)

Fig. 11 Visual representation method for RE: **a** Representation of relationships from Public DB; **b** Representation of relationships using statistical approach; **c** Representation of relationships using BERT approach

Result

Experimental setups

Table 2 shows the details of the system environment used for the experiments in this paper.

The split of annotations used for training, validation, and testing on the dataset used to fine-tune the NER and RE tasks with BioBERT is shown in Table 3.

The main hyper-parameters of the model utilized in GPDMiner, fine-tuned for the NER and RE tasks using these datasets, are shown in Table 4.

Experimental results

For the evaluation of our system, we performed a feature comparison with other text mining tools. One of the tools we compared against is the GENIA Tagger [33], a versatile tool encompassing morphological tagging, shallow parsing, and named entity recognition specialized for biomedical text. It has been meticulously trained on a diverse set of corpora, including the Wall Street Journal corpus, GENIA corpus, PennBioIE corpus, and NLPBA dataset for named entity recognition. Another noteworthy contender is tmTool [34], a web-based text mining utility equipped with batch processing capabilities. It excels in processing raw text from a wide spectrum of biomedical sources, including literature, patents, and medical records. tmTool integrates several entity tagging systems such as tmChem [35], DNorm [36], GNormPlus [37], and tmVar [38], enhancing its adaptability. We also examined PubTerm [39], a web-based tool tailored to support

Table 2 Experiment environment

System environment	
CPU	Intel® Core™ i9-10920X 12-core Processor
VGA	NVIDIA GeForce RTX 3090 24GB
OS	Windows 10

Table 3 Distribution of annotations

Dataset	Train	Dev	Test
NCBI disease	5130	787	960
BC2GM	5445	3060	4765
GAD	4960	0	534

Table 4 Hyper parameters

Parameter	Value of NER	Value of RE
Epochs	30	10
Optimizer	AdamW	AdamW
Max. length	192	128
Batch size	32	32
Learning rate	0.00005	0.00002

the analysis, annotation, and curation of biomedical scientific literature-a task integral to biomedical research, database curation, and clinical practice. EzTag, a web-based annotation tool, was another contender in our comparative evaluation. It empowers curators to annotate various biological concepts using pre-trained annotators like GNormPlus [37], tmVar [38], and TaggerOne [40]. Notably, it offers the flexibility to perform annotation tasks with or without existing training data, as well as the manual generation of training data. Lastly, TeamTat, a web-based annotation platform, offers efficient project management capabilities for collaborative annotation projects. It prioritizes project management features, facilitating multi-user annotation, figure display, and overall project coordination.

To gauge the performance of our NER and RE models, we employed standard metrics, namely precision, recall, and F1 score. In Table 5, you can find a comparative analysis of our fine-tuned NER model, tailored to our experimental environment, alongside NER models utilized in other systems. This evaluation provides insights into the efficacy of our system in comparison to established alternatives.

Experiments with the fine-tuned model in the gene/protein category show the best performance, with precision around 89%, recall around 93%, and F1-score around 91%. This compares favorably with other studies. In particular, other text mining tools such as ezTag, tmTool, and PubTerm use the GNormPlus model, which performs about 79% in terms of F1-score. In the disease category, we also experimented with BioBERT, which performs well among competing models with an F1-score of about 86%. Here, ezTag uses the TaggerOne model, while the tmTool and PubTerm tools use the DNorm model. The F1-scores of the TaggerOne and DNorm models are about 83% and 87%, respectively, which shows that our fine-tuned model is quite competitive in disease recognition.

For the RE task, the experimental results of the fine-tuned model show a precision of about 76.65%, a recall of about 91.1%, and an F1-score of 83.25%. There are no other text mining tools that provide RE results, making it difficult to compare performance on RE, which suggests that our system is unique for the RE task.

As shown in Table 6, both our system and other text mining tools provide built-in NER capabilities for disease and gene research domains. Unlike other text mining tools, our

Table 5 Comparison of NER models for genes/proteins, diseases

Entity types	Models	Precision	Recall	F1-score
Disease	Ours	0.8933	0.9300	0.9112
	TaggerOne Joint [40]	0.8510	0.8080	0.8290
	TaggerOne NER-only [40]	0.8350	0.7960	0.8150
	DNorm [36]	0.8030	0.7630	0.8720
	Sachan et al. [41]	0.8641	0.8831	0.8734
	CollaboNet [42]	0.8548	0.8727	0.8636
	LSTM-CRF (iii) of Habibi et al. [43]	0.8531	0.8358	0.8444
Gene/protein	Ours	0.8551	0.8899	0.8719
	GNormPlus [37]	0.7840	0.7920	0.7880
	Sachan et al. [41]	0.8181	0.8157	0.8169
	CollaboNet [42]	0.8049	0.7899	0.7973
	LSTM-CRF (iii) of Habibi et al. [43]	0.7750	0.7813	0.7782

Table 6 Function comparison with text mining tools

Tool	NER	Collaboration	RE
Ours	O	X	O
EzTag	O	X	X
TeamTat	O	O	X
GENIA tagger	O	X	X
TmTool	O	X	X
PubTerm	O	X	X

system has the unique ability to visualize relationship extraction results. This is a feature not found in other tools and is designed to help users intuitively understand and analyze complex relationships between recognized entities. This comparison gives us a clear picture of the strengths and weaknesses of each tool and can help you choose the best one for your specific research purposes.

Conclusions and future work

In this study, we introduce GPDMiner, an integrated data mining platform that innovatively supports the complex data analyses within the field of biomedicine. This system offers an amalgamation of diverse methodologies coupled with a user-centric interface, thereby mitigating the intricacies of data analysis while enhancing accuracy. A defining feature of GPDMiner is its integration of BERT-based NER and RE, along with a plethora of visualization functionalities. This allows for the intuitive comprehension of intricate analytical outputs. Advanced features for result storage and management facilitate efficient handling of large-scale data, further promoting research reproducibility and transparency. In response to the concerns raised in the review about the sufficiency of our datasets for fine-tuning our BERT-based model, we plan to develop a tagging tool system for augmenting our datasets through additional tagging. This will enable us to retrain our model with a more robust and comprehensive dataset, leading to significant performance improvements. The system we have currently implemented will serve as the base for this more advanced, final system that we plan to develop. The decision to not focus on performance analysis in this study is informed by the common observation that conventional text mining systems often exhibit significant performance degradation when deployed in real-world scenarios. Given GPDMiner's seamless operation in real-world settings without performance degradation, we find it unnecessary to conduct performance comparisons with other systems. Our primary objective is to provide a system that is easy to use and aids researchers and experts in their data analysis endeavors within the biomedical field. Future research directions will emphasize improvements in algorithmic performance and user experience optimization. By expanding the current functionalities and enhancing the intuitive nature of the interface, we aim to better support researchers and experts in more effectively analyzing and interpreting data. Through the continual integration of cutting-edge technologies and recent research findings, GPDMiner is poised to set new standards in data analysis within the realm of biomedicine. GPDMiner represents an important step forward for data mining in biomedical sciences, simplifying complex analytical processes and improving researchers' work efficiency. Future research is projected to broaden GPDMiner's functionalities and increase its efficacy to further drive its significance in research and decision-making in this area.

Acknowledgements
Not applicable.

Author contributions
Conceptualization, YJP and GJY; Methodology, YJP and GJY; Software, YJP and GJY; Validation, YJP, GJY, SJP, and CBS; Formal analysis, CBS; Investigation, YJP and GJY; Resources, SJP, and CBS; Data curation, YJP and GJY; Writing-original draft preparation, YJP and GJY; Writing—review and editing, YJP, GJY, SJP, and CBS; Visualization, YJP and GJY; Supervision, SJP and CBS; Project administration, YJP and GJY; Funding acquisition, SJP and CBS. All authors have read and agreed to the published version of the manuscript.

Funding
This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [24ZR1200, DNA Based National Intelligent Core Technology Development]. This work was supported by internal fund/grant of Electronics and Telecommunications Research Institute (ETRI) [23YB1600, Exploratory and Strategic Research of ETRI-KAIST ICT Future Technology]. The present Research has been conducted by the Research Grant of Kwangwoon University in 2023.

Availability of data and materials
All data generated or analyzed during this study are available upon reasonable request from the corresponding author.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 8 November 2023 Accepted: 19 February 2024
Published online: 06 March 2024

References

1. Börner K, Chen C, Boyack KW. Visualizing knowledge domains. Ann Rev Inf Sci Technol. 2003;37:179–255.
2. Adams J. The fourth age of research. Nature. 2013;497:557–60.
3. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science. 2002;298:1912–34.
4. Rajman M, Besançon R. Text mining: natural language techniques and text mining applications. In: Data Mining and Reverse Engineering: Searching for Semantics. IFIP TC2 WG2. 6 IFIP seventh conference on database semantics., 1998;50–64.
5. Feldman R, Sanger J. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge: Cambridge University Press; 2007.
6. Kulkarni S, Singh A, Ramakrishnan S. G. and Chakrabarti. Collective annotation of wikipedia entities in web text. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining., 2009;457–466.
7. Wei CH, Kao HY, Lu Z. Pubtator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res. 2013;41:518–22.
8. Hearst MA. Untangling text data mining. In: Proceedings of the 37th annual meeting of the association for computational linguistics on computational linguistics 1999;3–10.
9. Kao A, Poteet SR. Natural language processing and text mining. London: Springer; 2007.
10. Sang EF, De Meulder F. Introduction to the CoNLL-2003 Shared task: language-independent named entity recognition. In: Proceedings of CoNLL-2003., 2003;142–147.
11. Doddington GR, Mitchell A, Przybocki MA, Ramshaw LA, Strassel SM, Weischedel RM. The Automatic Content Extraction (ace) Program-tasks, Data, and Evaluation. In: Lrec., 2004;837–840.
12. Piskorski J, Pivovarova L, Šnajder J, Steinberger J, Yangarber R. The first cross-lingual challenge on recognition, normalization and matching of named entities in slavic languages. In: Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing. The Association for Computational Linguistics., 2017;76–85.
13. Segura-Bedmar I, Martínez Fernández P, Herrero Zazo M. Semeval-2013 Task 9: extraction of drug-drug interactions from biomedical texts. In: Association for Computational Linguistics., 2013;341–350.
14. Bossy R, Golik W, Ratkovic Z, Bessieres P, Nédellec C. BioNLP Shared Task 2013-an overview of the bacteria biotope task. In: Proceedings of the BioNLP Shared Task 2013 Workshop., 2013;161–169.

15. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc. 2011;18:552–6.
16. Balog K, Serdyukov P, De Vries AP. Overview of the TREC 2010 entity track. In: TREC., 2010;294.
17. Petasis G, Cucchiarelli A, Velardi P, Paliouras G, Karkaletsis V, Spyropoulos CD. Automatic adaptation of proper noun dictionaries through cooperation of machine learning and probabilistic methods. In: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval., 2000;128–135.
18. Nadeau D, Sekine S. A survey of named entity recognition and classification. Lingvisticae Investigationes. 2007;30:3–26.
19. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. In: Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: human language technologies., 2016;260–270.
20. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag. 1996;17:37–37.
21. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet. 2006;7:119–29.
22. Jang H, Lim J, Lim JH, Park SJ, Lee KC. BioProber: software system for biomedical relation discovery from PubMed. In: 2006 International conference of the IEEE engineering in medicine and biology society., pp. 2006;5779–5782.
23. Sung TT, Sohn CB, Park SJ. GDMiner: gene-disease relation miner system. In: Proceedings of the 12th international workshop on data and text mining in biomedical informatics (DTMBIO 2018) 2018.
24. Ruder S. Neural transfer learning for natural language processing. PhD Thesis, National University of Ireland, Galway. Galway 2019.
25. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J Royal Soci Interface. 2018;15:20170387.
26. Berger AL, Pietra VJD, Pietra SAD. A maximum entropy approach to natural language processing. J R Soc Interface. 1996;22:39–71.
27. Devlin J, Chang M-W, Lee K. Toutanova. K. BERT: pre-training of Deep Bidirectional Transformers for Language Understanding; 2019.
28. Dogan RI, Leaman R, Lu Z. Ncbi disease corpus: a resource for disease name recognition and concept normalization. J Biomed Inform. 2014;47:1–10.
29. Smith L, Tanabe LK, Ando RJN, Kuo CJ, Chung IF, Hsu CN, et al. Overview of biocreative II gene mention recognition. Genome Biol. 2008;9:1–19.
30. Bravo A, Piñero J, Queralt-Rosinach N, Rautschka M, Furlong LI. implications for translational research. Extraction of relations between genes and diseases from text and large-scale data analysis. BMC Bioinform. 2015;16:1–17.
31. Heer J, Shneiderman B. Interactive dynamics for visual analysis: a taxonomy of tools that support the fluent and flexible use of visualizations. Queue. 2012;10:30–55.
32. Plaisant C, Shneiderman B. Show Me! Guidelines for producing recorded demonstrations. In: 2005 IEEE symposium on visual languages and human-centric computing (VL/HCC'05)., 2005;171–178.
33. Tsuruoka Y. GENIA Tagger. http://www.nactem.ac.uk/GENIA/tagger/ 2009.
34. Wei CH, Leaman R, Lu Z. Beyond accuracy: creating interoperable and scalable text-mining web services. Bioinformatics. 2016;32:1907–10.
35. Robert L, Wei C-H, Lu Z. tmchem: a high performance approach for chemical named entity recognition and normalization. J Cheminform. 2015;7(1):1–10.
36. Robert L, Doğan LI, Lu Z. Dnorm: disease name normalization with pairwise learning to rank. Bioinformatics. 2013;29:2909–17.
37. Wei C-H, Kao H-Y, Lu Z. Gnormplus: an integrative approach for tagging genes, gene families, and protein domains. BioMed Res Int. 2015. https://doi.org/10.1155/2015/918710.
38. Wei CH, Phan L, Feltz J, Maiti R, Hefferon T, Lu Z. tmvar 20: integrating genomic variant information from literature with dbsnp and clinvar for precision medicine. Queue. 2018;34(1):80–7.
39. Garcia-Pelaez J, Rodriguez D, Medina-Molina R, Garcia-Rivas G, Jerjes-Sánchez C, Trevino V. Pubterm: a web tool for organizing, annotating and curating genes, diseases, molecules and other concepts from pubmed records. Database. 2019.
40. Leaman R, Lu Z. Taggerone: joint named entity recognition and normalization with semi-markov models. Bioinformatics. 2016;32(18):2839–46.
41. Sachan DS, Xie P, Sachan M, Xing EP. Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition. In: Machine learning for healthcare conference., 2018;383–402.
42. Yoon W, So CH, Lee J, Kang J. Collabonet: collaboration of deep neural networks for biomedical named entity recognition. BMC Bioinform. 2019;20(10):55–65.
43. Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics. 2017;33(14):37–48.

Publisher's Note