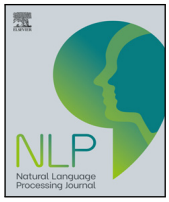
Natural Language Processing Journal 8 (2024) 100095

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp



Scholarly recommendation system for NIH funded grants based on biomedical word embedding models

Zitong Zhang *, Ashraf Yaseen, Hulin Wu

Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, 77030, TX, USA

ARTICLE INFO

Dataset link: https://reporter.nih.gov/exporter/projects, https://pubmed.ncbi.nlm.nih.gov/download/

Keywords:
NIH
Funded grant
Word embedding
Recommendation system

ABSTRACT

Objective: Research grants, which are available from several sources, are essential for scholars to sustain a good standing in academia. Although securing grant funds for research is very competitive, being able to locate and find previously funded grants and projects that are relevant to researchers' interests would be very helpful. In this work, we developed a funded-grants/projects recommendation system for the National Institute of Health (NIH) grants.

Methods: Our system aims to recommend funded grants to researchers based on their publications or input keywords. By extracting summary information from funded grants and their associated applications, we employed two embedding models for biomedical words and sentences (biowordvec and biosentvec), and compare multiple recommendation methods to recommend the most relevant funded grants for researchers' input

Results: Compared to a baseline method, the recommendation system based on biomedical word embedding models provided higher performance. The system also received an average rate of 3.53 out of 5, based on the relevancy evaluation results from biomedical researchers.

Conclusion: Both internal and external evaluation results prove the effectiveness of our recommendation system. The system would be helpful for biomedical researchers to locate and find previously funded grants related to their interests.

1. Introduction

Researchers must have the ability to acquire grants in order to succeed academically. With plentiful resources from funding agencies such as the National Institute of Health (NIH), the National Science Foundation (NSF) and others, researchers are presented with many potential opportunities. However, locating and assessing the viability of grants can be a tricky and lengthy task. To better equip applicants, records of funded grants are available from government and private sources, though the sheer magnitude and disorderly organization of the databases often prolongs the search.

As the largest public funder of biomedical research in the world, NIH invests more than 32 billion dollars a year to fund research which led to breakthroughs and new treatments, healthier lives, and build the research foundation that drives discovery (U.S Department of Health and Human Services, 2023). Each year, thousands of funded grants would be published in their database, and millions of publications and applications related to these projects would be generated. As such, a recommendation system which connects the interests of researchers and these funded grant resources would greatly increase the efficiency of grant application and help the researchers with their applications.

In this paper, we describe the development of a content-based grant recommendation system that recommends funded grants to researchers as a ranking problem of grants matched against information extracted from publications or keywords provided by researchers. A comparison between basic IR-based system (BM25 & TF-IDF) and our approach in employing the sentence embedding models for biomedical words: Biowordvec & Biosentvec (Zhang et al., 2019; Chen et al., 2019) was conducted and detailed in this work. The recommender is evaluated on both precision and user-rating, and is proved to be an efficient recommendation system for researchers.

The work described herein is part of the Virtual Research Assistant (VRA) platform developed by our team at The University of Texas Health Science Center at Houston. VRA platform includes several scholarly recommenders for papers, public datasets, collaborators, grant announcements, journals and conferences (Patra et al., 2020a; Zhu et al., 2023; Patra et al., 2020b).

1.1. Related works

The most commonly used approaches in recommendation systems include content-based filtering (CBF) and collaborative filtering (CF).

* Corresponding author.
E-mail addresses: Zitong.Zhang@uth.tmc.edu (Z. Zhang), Ashraf.Yaseen@uth.tmc.edu (A. Yaseen), Hulin.Wu@uth.tmc.edu (H. Wu).

https://doi.org/10.1016/j.nlp.2024.100095
Received 31 January 2024; Received in revised form 9 July 2024; Accepted 2 August 2024

2949-7191/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Because of the requirement to handle contextual information, assess keywords, and explore academic resource topics, the majority of scholarly recommendation systems have been constructed using CBF. Developing a content-based profile for researchers typically focuses on the user's preference model and their interactions with the recommendation system, utilizing a weighted vector of item features. For instance, Hong et al. (2012) devised a literature recommendation approach that builds a user profile using extracted keywords. They employed cosine similarity to gauge the similarity between a given topic and collected papers, thereby suggesting initial list of publications as recommendations for each topic.

To consider the popularity and rating of recommended items, researchers might use collaborative filtering (CF). For example, Tang and McCalla constructed user profiles via a co-author network to build a serendipitous paper recommendation based on scholarly social networks (Tang and McCalla, 2009). However, the cold start problem, which represents the need for a large volume of existing data for a collaborative filtering system to make accurate recommendations, has limited the application of collaborative filtering in scholarly recommendation systems (Lika et al., 2014).

1.1.1. History of scholarly recommendation systems

Scholarly recommendation system is an important tool for identifying prior and related resources such as literature, collaborators, conferences and journals, and more. In 1998, Bollacker et al. published the initial type of scholarly recommender system, which aimed to recommend publications using content-based similarity methods (Bollacker et al., 1998). Since then, recommendation systems have been widely used by researchers in various scholarly fields.

In the area of recommending collaborators, numerous applications are published;. By representing queries and documents as vectors in a multi-dimensional space, these vectors can be utilized to calculate the relevance or similarity among collaborators. For example, in 2012, Gollapalli et al. addressed a scholarly content-based researcher recommendation system by computing the similarity between researchers based on their personal profiles extracted from their publications and academic homepage (Gollapalli et al., 2012). Also, based on this co-authorship network transformed from researchers' publication activities, several methods for link prediction and edge weighting, such as Benchettara's topological dyadic supervised machine learning approach in solving the problem of link prediction in co-authoring networks, were utilized (Benchettara et al., 2010).

Recommendation systems can also help reduce the cognitive burden that comes with selecting the appropriate conference or journal for publishing work, through academic venue recommendation systems for conferences and journals. The most popular recommendation approach for venue was based on generating and analyzing a variety of networks using different types of metadata, including citations, co-authorship, references, social proximity, etc.

To recommend conferences to users, Asabere and Acakpovi developed a user-based social context-aware filter using Breadth First Search and Depth First Search on a knowledge graph created by computing the Social Ties between users, and added geographical, computing, social, and time contexts (Asabere and Acakpovi, 2019; Asabere et al., 2018).

In many cases, the authors may encounter difficulties in finding the suitable journals for their manuscripts. A journal recommendation system can alleviate the authors' burden by selecting appropriate journals to publish, and it may also reduce the burden of the editors by rejecting manuscripts that do not align with the journals' scopes. By utilizing the similarity between the provided keywords and journal keywords to identify suitable journals, lots of recommendation applications, such as eTBLAST and SJFinder, were developed and published (Errami et al., 2007; Anon, 2014).

Since 2021, our team has created several scholarly recommendation systems for datasets, collaborators and grant announcements. Using datasets information including title, abstract and summary, Patra et al. applied content-based recommendation methods on Gene Expression Omnibus (GEO) data and proved the methods could help increase data reusability (Patra et al., 2020a,c). Combining GEO repository with other biomedical data resources, Zhang et al. experimented with information retrieval paradigms (BM25, TF-IDF, etc.) for dataset recommendation to researchers based on metadata extracted (Zhang and Yaseen, 2023). In 2020, Zhu et al. implemented graph neural networks for collaborator recommendation by capturing the complex relationship and dynamic dependencies between researchers (Zhu and Yaseen, 2022). Furthermore, Zhu et al. (2021) developed a recommender of scholarly papers for researchers, based on publicly available database using a BERT-based model (Zhu et al., 2021). Based on BERT-based datasets recommendation system, this group also performed a sensitivity analysis on the training class imbalance (Zhu et al., 2022).

1.1.2. Research in grant recommendation systems

To the best of our knowledge, there is no published recommendation system for funded grant projects. However, several studies that focus on grant announcements have been published, and their findings have been proven to be efficient, including our work published in Zhang et al. (2023).

In 2023, using researchers' publications to build personal profiles, Zhu et al. proposed a grant announcements recommendation system for NIH grants (Rajaraman and Ullman, 2011). This is the latest publication of the few studies available for recommending grant funding announcements since 2015, when Kamada et al. developed a search engine for finding Japanese research announcements (Kamada et al., 2015, 2016). These publications describe a keyword-based search engine using TF-IDF and association rules and apply the searching of funding announcements of Japan.

In this research, we developed our funded grant project recommendation system using content-based recommendation methods.

2. Material and methods

Our recommender utilizes the sentence embedding models for biomedical words: Biowordvec & Biosentvec. For comparison purpose, basic IR-based systems using BM25 and TF-IDF were used. Input to the system can be an abstract of a paper, description of a research idea, or just keywords. The recommender will then generate a ranked list of funded grants' descriptions as recommendations.

2.1. Data collection

Information about NIH funded grants were extracted from ExPORTER (ExPORTER, 2024), and information about related publications were extracted from PubMed (PubMED, 2023), a free search engine of references and abstracts on life sciences and biomedical topics accessing the MEDLINE database (MEDLINE, 2021). Data collection methods and summaries of data are described next.

ExPORTER is an important component of the NIH "open government" initiatives, which aims at providing more transparency into NIH activities, and improving the reusability and quality of data collected. Based on RePORTER (NIH RePORTER, 2024), a publicly available electronic searching platform for a database of research projects, accessing publications and patents resulting from NIH funding, ExPORTER provides a large quantity of administrative data existed.

The database from ExPORTER consists of four parts. From several extant databases where NIH funded projects have been cited, ExPORTER draw information about these grants. For each funded grant, there is a core project associated with one or more applications. The first part of ExPORTER database contains the information about these core projects including their IDs, project titles, terms, principle investigators, the association between projects and their related applications, and other clinical trial information (RePORTER_PRJ_C_FY2020.csv). The second part contains the basic information about these grant applications, including application IDs and abstracts (RePORTER_PRJABS_C_FY2020).

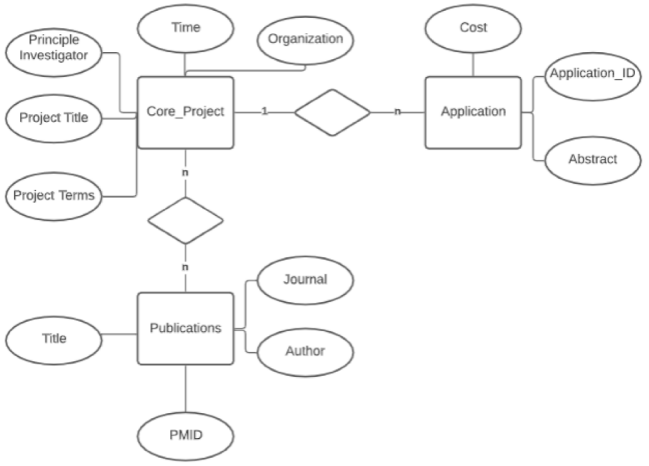


Fig. 1. Linkage relationship between funded grants, applications and publications.

After funds are granted to investigators, grant projects normally produce multiple publications. Via the NIH internal tool SPIRES (National Institution of Health, 2012), ExPORTER integrates publications and gathers information from NCBI My Bibliography and the NIH Intramural Database to link NIH projects with research publications. While the third part of database lists all the IDs of biomedical publications on PubMed in 2020 (REPORTER_PUB_C_2020), the fourth part of database contains the association between the projects and the publications (REPORTER_PUBLNK_C_2020).

Fig. 1 shows the linkage relationship between these funded grant projects, applications and publications, and the key features we extracted from the database for these resources.

For our experiment detailed in this work, we considered all grant projects that were founded by NIH in the year 2020 (N = 63,534 projects). We then selected projects starting with "R", which represents "research project". As a result, 43,806 applications and 43,793 full projects were extracted from the ExPORTER database as raw dataset. 38,196 of these projects are related to only 1 application, while 25,40 are related to 2 applications, 162 are related to 3, and only 11 of them are related to more than 4 applications. ExPORTER database listed 112,371 publications citing support from these projects. For the evaluation of recommendation results, 73,913 linkage relationship between these projects and publications within 2020 are then extracted for the evaluation of recommendation results.

NIH projects funded in the fiscal year of 2008 and beyond, the research, condition, and/or disease category in which a project falls is determined by an automated text-mining tool (Anon, 2023). The tool would derive concepts by mining the text information about the projects, including its' title, abstract, specific aims, and the public health relevance stated by the investigators. These derived concepts would be concluded as project terms. We then used this as a filter in subsequent experiment to generate specialized recommendations.

2.2. Data processing

Fig. 2 describes the data processing pipeline of our funded grants recommendation system.

First, the databases of funded grants were treated as text objects where the text of a funded grant includes its project title, project terms, and abstract of applications. To make recommendations based on publications, we also extract text object from database of publications based on their title and abstract. Pre-processing was performed on both input text information and funded grants information by removing the stop words, punctuation, links, and other junk words with low value. Subsequently, the NLTK WordNet lemmatizer would be used to get the root forms of the words (Bird et al., 2009).

In the following section of this paper, we evaluated the performance of our funded grant recommendation system on datasets with and without filtering. To recommend grant projects based on filtered research fields, the database would be divided into subgroups based on the Administering institute or center marked in core project IDs, and recommendations would be made between those filtered funded grants and their relevant publications.

To improve the precision of the recommendation results, we applied 2 different word embedding methods to process the lemmatized textual data. Then the cosine similarity between the input and the funded grants are calculated based on text vectors. For each researcher using this recommendation system by inputting keywords, paragraphs, or abstracts of their publications, a list of recommended funded grants will be generated based on the rank of similarity score between the input and the information of funded grants.

3. Theory

To provide precise recommendation, we tested multiple combinations of attributes in the database funded grant projects. Based on our former experiments representing funded grants with keywords only, the performance of recommendation systems is not good as we expected. Thus, the core project title, terms, and the abstracts of its associated applications are all considered as the key features which describe the funded grants.

We applied two Information Retrieval (IR)-based models, BM25 and TF-IDF, as baseline systems, and two proposed pre-trained word-embedding models based on biomedical text, Biowordvec and Biosentvec, to vectorize these text objects.

3.1. Baseline: IR-based

The most similar funded grants can be recommended based on input text simply by comparing the cosine similarity of the input text and grant vectors using (1):

$$sim(r,d) = \cos\left(v_r, v_d\right) = \frac{v_r \cdot v_d}{\|v_r\| \cdot \|v_d\|} \quad (1)$$

In Eq. (1), d represents all the grants that can be recommended to the original text r, and $\cos(v_r, v_d)$ is the cosine similarity between text vector (v_r) and grants vector (v_d).

Two baseline recommenders were implemented using the Information Retrieval (IR) based methods: Term Frequency-Inverse Document frequency (TF-IDF) and BM25.

3.1.1. BM25

BM (best matching) is a probabilistic retrieval ranking function using the terms appearing in each document. We utilize Okapi BM25 for matching vectors generated from each grant (Robertson and Zaragoza, 2009).

For each funded grant, the title, keywords, and application abstract information were pre-processed and normalized and then converted into a single vector. BM25 generates similarity scores between input text information and the grant vectors for the recommendation.

3.1.2. TF-IDF

TF-IDF (term frequency – inverse document frequency) is a statistical representation of the importance of a word or term, among the whole collection of documents (Rajaraman and Ullman, 2011). If in a given document, a given term reached a high term frequency, and a low document frequency of the term is achieved within the whole collection of documents, there would be a high weight for this term. Hence, common terms that may bias the recommendation would be filtered out based on the weights.

The text objects for the inputs and all projects were preprocessed and normalized for the conversion. Then the lemmatized text would be converted into single text vectors. Finally, each input text vector would be compared with grant vectors to generate the recommendation score.

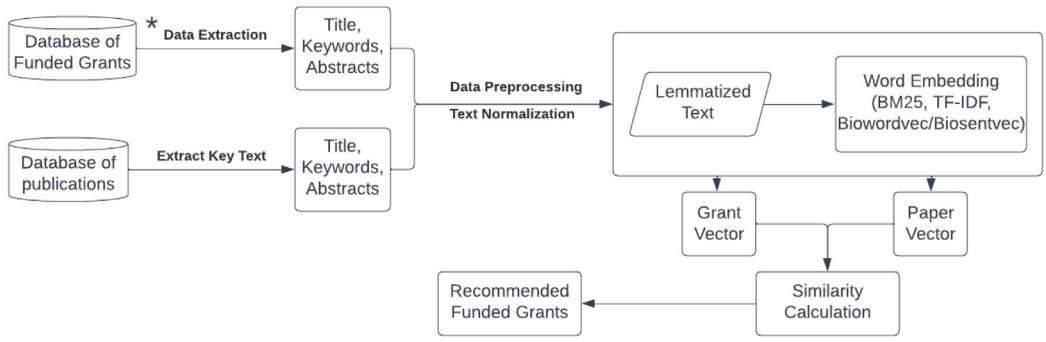


Fig. 2. Pipeline for funded grant recommendation system.
* For recommendations based on filtered grant projects, we could apply filtering at this point based on research fields marked in core project IDs.

3.2. Proposed method: Embedding methods for biomedical words and sentences

While using the baseline IR system for the recommendation, we noticed the differences between the language level of application abstracts and project terms. The language level for application abstract tends to be specific, while the project terms and keywords often use more generalized descriptions. The issue of biomedical words out of vocabulary also influences precision. To overcome these problems, we proposed 2 biomedical word embeddings: Biowordvec using fastText embedding model, and Biosentvec using sent2vec (Zhang et al., 2019; Chen et al., 2019).

3.2.1. Biowordvec: Biomedical word embeddings with fastText

BioWordVec is a set of word embeddings using the subword embedding Model, it is trained on two public available data resources: biomedical publications in PubMed and domain knowledge in MeSH data.

The Biowordvec method consist of 2 steps: First, constructing a MeSH term graph based on the MeSH RDF data, and generating a list of MeSH term sequences using random sampling strategy; Then these term sequences could be learned in a unified n-gram embedding space, using the subword embedding model.

A proposed network embedding methods using skip-gram models have been applied to convert network into nodes sequences (Perozzi et al., 2014; Tang et al., 2015). In Biowordvec, the heading nodes which represent the relationship among the MeSH term graph would be converted into ordered sequences. These ordered heading sequences would then be combined with PubMed sentence sequences to train word embeddings.

Based on the skip-gram model, Bojanowski et al. proposed a subword embedding model named fastText (Bojanowski et al., 2017). Using an unlabeled dataset where represented each word as a sum of its n-grams vector representation, the model learned character n-grams distributed embedding. Compared to the word2vec model (Mikolov et al., 2013), the subword embedding model can effectively gain information from the internal word structure, and improve the embedding quality. For example, many specialized compound words in real-world biomedical projects, such as "gammaproteobacteria", are rarely existed, or even out of vocabulary words in the training dataset. Thus, it would be difficult for the classic word2vec model to understand and train these compound words. To deal with such problems, the subword embedding model would be more suitable. For example, the subword embedding model would be able to learn the distributed representations of all character n-grams of the compound word "gammaproteobacteria" by dividing the compound word into three commonly used sub words, "gamma", "proteo" and "bacteria", in training dataset, and effectively integrate the subword vectors to create the final embedding for the compound word without losing information. This subword embedding model was applied in Biowordvec model to train word embeddings based on the joint text sequences of MeSH and PubMed.

The fastText subword embedding model is originated from the continuous skip-gram model. The objective function of the skip-gram model for given word sequence w1, w2, …, w_T, is defined as follow:

$$J = max \frac{1}{T} \sum_{1}^{T} \sum_{c \in C_t} log p(w_c | w_t)$$

where C_t represents the set of the surrounding words of wt. The probability $p(w_c|w_t)$ represents the probability of observing surrounding word w_c given current word w_t:

$$p\left(w_c | w_t\right) = \frac{e^{s(w_t, w_c)}}{\sum_{j=1}^{W} e^{s(w_t, w_j)}}$$

while $s(w_t, w_c)$ is the scoring function. The scoring function,$s(w_t, w_c) = u_{wt}^{t} v_{wc}$, is defined by the original skip-gram as scalar product, where u_{wt} and v_{wc} denotes the vectors of two words w_t and w_c. Using this function, the original skip-gram model would not be able to utilize the subword information, since only one distinct vector could be generated from each word.

To solve this problem, the fastText model considered each word as a list of character n-grams. For example, the word "notation" will be represented by the character 4 grams including <#not, nota, otat, tati, atio, tion, ion#> and the word itself <notation>. Compared to the original skip-gram model, the proposed subword embedding model improve the representation by defining the scoring function, s(wt, wc) as the $\Sigma_{g \epsilon (1,\ldots,G)} z_g^T v_c$. In this definition, (1, …, G) is the n-gram set of wt, Z_g represents the vector of character n-gram g, and v_c is the vector of word w_c. Hence, each individual word would be represented as the sum of the vector representations of its n-grams. And the fastText model would hence learn the distributed representation of character n-grams. By sharing these representations of n-grams across words, the fastText model significantly improves its efficiency in learning reliable embedding for the out of vocabulary words.

In the pre-trained model, the joint text sequences from MeSH and PubMed are used to as the input. For the MeSH term sequences, the objective function is defined as:

$$J_{MeSH} = \frac{1}{N} \sum_{1}^{N} \sum_{c \in C_t} log p(D_c | D_t)$$

where N represents the total number of MeSH main headings.

For the PubMed publications, the objective function is defined as:

$$J_{PubMed} = \frac{1}{T} \sum_{1}^{T} \sum_{c \in C_t} log p(w_c | w_t)$$

where T represents the total number of vocabulary.

Finally, these two objective functions were linearly combined and maximized to train the model based on MeSH term sequences and PubMed text sequences jointly:

$$J = J_{PubMed} + J_{MeSH}$$

Table 1
Selective parameters for the models used in the experiments.

	Method	Parameters
Baseline	BM25	b = 0.75, k_1 = 1.5
	TF-IDF	Ngram $_{range}$ = (1, 2), min $_{df}$ = 2, max_features = 2000
Proposed	Biowordvec	word vector dimension = 200, negative sample size = 10
	Biosentvec	negative samples = 10, t = 5 * 10^{-6}, max length of word ngram = 2

Since the MeSH term sequences and PubMed sentence sequences share multiple biomedical phrase and subwords, the subword embedding model would combine the n-gram representations between MeSH and PubMed, integrate them into one single unified embedding space, and train the two objective functions together using the fastText subword embedding model.

3.2.2. Biosentvec: Biomedical sentence embeddings with sent2vec

BioSentVec is created by applying sent2vec, an advanced unsupervised model, to both biological and clinical texts at a large scale (Chen et al., 2019). It is the first public available sentence embeddings trained from both scholarly publications in PubMed and clinical notes in the MIMIC-III Clinical Database.

Sent2Vec is an unsupervised model which trains both composition and the embedding vectors themselves. It composes sentence embeddings by using word vectors along with n-gram embeddings. Literally, these models are all originate from the form:

$$\min_{U,V} \sum_{S \in C} f_S(UV\iota_s)$$

for two parameter matrices U ∈ $R^{k\times h}$ and V ∈ $R^{h\times|V'|}$. Where the columns of matrix U denote the target word vectors, V' represents the vocabulary, and the columns of V denote the source word vectors already learned. For each random-length given sentence S, a binary vector $\iota S \in \{0,1\}^{|V|}$ would be encoded, defined as the indicator vector.

Theoretically, the Sent2vec model can be interpreted as a natural extension of the word-contexts from C-BOW to a larger sentence context (Mikolov et al., 2013). By using the unsupervised objective function, the words and phrases would be specifically optimized to gain additional combination over the sentence.

A context embedding v_w and target embedding u_w for each word w would be generated using embedding dimension h and k = |V| . By learning source embeddings n-grams within each sentence, and averaging the n-gram embeddings along with its component parts, sentence embedding model achieves better performance. As an average of its component words' embedding, the sentence embedding v_S for S is modeled as equation:

$$v_s := \frac{1}{|R(S)|} V\iota_{R(S)} = \frac{1}{|R(S)|} \sum_{w \in R(S)} v_w$$

where R(S) is the list of n-grams present in sentence S (including unigrams). The outputs would then be approximated by negative sampling to predict missing words from the text (Mikolov et al., 2013). Given the binary logistic loss function l: $x \rightarrow \log(1 + e^{-x})$ coupled with negative sampling, the formula of unsupervised training objective is defined as:

$$\min_{U,V} \sum_{S \in C} \sum_{w_t \in S} (\mathrm{l}\left(u_{w_t}^T v_{S\setminus\{w_t\}}\right) + \sum_{w' \in N_{w_t}} \mathrm{l}\left(-u_{w'}^T v_{S\setminus\{w_t\}}\right))$$

In this case, S means the current sentence and N_{wt} means the set of words sampled negatively for the word wt ∈ S. The negatives are sampled following a multinomial distribution, where each word w is associated with a probability:

$$q_n(w) := \sqrt{f_w} / (\sum_{w_i \in v} \sqrt{f_{w_i}})$$

and f_w represents the normalized frequency of w in the dataset.

Based on Lapata et al. (2017), to select the possible target unigrams (positives), each word w was discarded with probability–1 - q_p(w) defined as:

$$q_p(w) := \min\left\{1, \sqrt{\frac{t}{f_w}} + \frac{t}{f_w}\right\},$$

where variable t is the subsampling hyper-parameter. Subsampling reduces the bias caused by frequently appeared words which strongly influence the training progress in the prediction task. As a combination of both positive and negative subsampling distribution, the training objective function for sent2vec is defined as:

$$\min_{U,V} \sum_{S \in C} \sum_{w_t \in S} (q_p(w_t) \mathrm{l}(u_{w_t}^T v_{S\setminus\{w_t\}}) + \left|N_{w_t}\right| \sum_{w' \in v} q_n(w') \mathrm{l}(-u_{w'}^T v_{S\setminus\{w_t\}})).$$

Using the sent2vec model, BioSentVec embeddings are trained based on both clinical notes from MIMIC-III Clinical Database and the PubMed.

3.2.3. System parameters

Parameters used for the baseline and our proposed methods are listed in Table 1. Based on Mikolov et al. (2013), all n-grams (3 <= n <= 6) were extracted by the subword model for training word representations in Biowordvec embeddings. For the Biosentvec embeddings, L2 regularization was applied to the sent2vec model.

3.3. Evaluation metrics

To measure the accuracy of different algorithms for grant recommendation, we calculated the metrics against the reference relationship between funded grants and PubMed publications in 2020. Metrics include Precision at position n (P@n), which is used to express the proportion of recommendations that are relevant; and Recall at position n (Recall@n), which is used to present the proportion of relevant items found in the top-n recommendations.

Metrics were calculated against the ground truth association between the input publications and the recommended funded grant project lists. As mentioned in Section 2.1 above, we extracted the many-to-many relationships between grant projects and publications citing support from these projects as link tables. If a publication is cited by a NIH research project, their relationship will appear as an association in the link table. While we use publications related to NIH projects in 2020 as inputs, we could figure out whether the recommended projects are relevant to the input publication or not based on these relationships.

Based on the relationship network between funded grants and publications, we generated 3 rating criteria for the relevancy. The ratings are:

- No relevance: There is no relationship between the publication and the recommended funded grant projects.
- Partial relevance: The recommended projects can be linked to the input publication according to the reference relationship network. For example, if paper 1 is referenced by project A and project B, and paper 2 is referenced by project B and project C, we would say that paper 1 and project C is partially related, via their relationship between project B and paper 2.
- Strict relevance: The recommended projects reference the publication or is referenced by the input publication.

In order to precisely define the evaluation metrics Recall@n and Precision@n, we supplement the confusion matrix as shown below in Table 2.

- Recall@n: At the n^{th} recommended item, this metric shows the proportion of relevant items that are recommended.

$$Recall@n = \frac{TP@n}{Total\ \ Relevant}$$

Table 2
Confusion Matrix for Evaluation Metrics.

	Recommended	Not recommended	Total
Relevant	True positive (TP)	False negative (FN)	Total relevant
Not Relevant	False positive (FP)	True negative (TN)	Total Not relevant
Total	Total recommended	Total Not recommended	Overall Total

Table 3
Internal evaluation results for general datasets.

	P@100 (Partial)	P@100 (Strict)	Recall @100
Baseline (IR)			
BM25	0.613	0.504	0.445
TF-IDF	0.590	0.481	0.421
Proposed			
Biowordvec	0.743	0.629	0.510
Biosentvec	0.759	0.632	0.513

Table 4
Datasets for Recommendation Based on Subgroups.

Subgroup	Eye	Mental health	Cancer	Total
Publications	1871	3231	6612	11 714
Related Projects	1394	2461	5020	8875

- Precision@n: At the nth recommended item, this metric shows the proportion of the recommended items that are relevant.

$$Precision@n = \frac{TP@n}{Total\ Recommended}$$

4. Results

4.1. Internal evaluation

4.1.1. Recommendations based on general dataset

Using the related publications for NIH research project grants in 2020, we recommend grants based on text information extracted from the publications and evaluate the precision of the recommendation based on the relationship between publications and grants. 13,796 randomly selected publications and their related funded grant projects are used as evaluation samples in this experiment. Table 3 shows the results for different metrics.

The evaluation metrics are calculated for the methods we used as shown in Table 3. Methods with biomedical word embedding model significantly increase the precision and recall of the recommendation, which proves the higher efficiency of our proposed methods. Among the proposed models using biomedical word embedding methods, the performance between the two models were similar, while the Biosentvec model showed slightly better performance.

4.1.2. Recommendation based on filtered grants

In order to make precise recommendations, we provided recommendations based on datasets and grant projects and publications filtered by their Administering Institute or Center.

A two-character code is used to designate the agency, NIH Institute, or Center administering the grant in NIH funded grants ID. The definition of these institution and center code is provided in National Institution of Health (2024). Filtering the 43,807 NIH research projects in 2020 we extracted, 34 subgroups were obtained based on their research fields, such as general medicine (GM), allergy and infectious disease (AI), cancer (CA), heart and lung and blood (HL), etc. Fig. 3 describes the distribution of these subgroups.

In this experiment, 11,714 publications and their related funded grant projects randomly selected from 3 different subgroups are used as evaluation samples. The detailed information about these subgroups is shown in Table 4.

As shown in Table 5, compared to recommendation based on general datasets without filtering, recommendation based on grants filtered by research fields significantly increase the recall value of the models. Also, for most of the subgroups, higher strict precision values are observed compared to the general datasets without filtering. Both comparisons indicate that recommendation among filtered groups would provide funded grants recommendation with higher performance.

Table 5
Internal evaluation results for datasets filtered by research fields.

	P@100 (Partial)	P@100 (Strict)	Recall@100 (Strict)
Eye (N = 1871)			
BM25	0.660	0.568	0.530
TF-IDF	0.645	0.551	0.510
Biowordvec	0.769	0.665	0.644
Biosentvec	0.782	0.681	0.665
Mental Health (N = 3231)			
BM25	0.670	0.559	0.562
TF-IDF	0.647	0.537	0.536
Biowordvec	0.791	0.704	0.727
Biosentvec	0.803	0.721	0.748
Cancer (N = 6612)			
BM25	0.593	0.515	0.478
TF-IDF	0.567	0.492	0.452
Biowordvec	0.721	0.641	0.619
Biosentvec	0.738	0.659	0.639

Table 6
External evaluation results for top 10 recommended projects.

Researcher ID	Average rating score (1–5)	Highest rating score (1–5)
Researcher 1	4	5
Researcher 2	3.38	5
Researcher 3	3.75	5
Researcher 4	3.71	5
Researcher 5	3.25	5
Researcher 6	3.63	5
Researcher 7	3	4
Average	3.53	5

4.2. External evaluation

To test the efficiency of this funded grant recommendation method as an application, we invited several faculty members in our school to evaluate this system. We asked them to provide an abstract of a paper they had published or a research-summary paragraph as input and match their input to the metadata extracted from the database of NIH funded research projects.

For each faculty member, we recommended 10 funded grant projects based on the input they provided or the abstract of their publications. The participants then rated these 10 recommendations on a scale from 1 to 5, 5 being highest. The rating criterion is the relevancy between the recommended projects and the researchers' interest in this topic. 5 is defined as "most relevant", which means this recommended grant is the most relevant and the user is interested in this project. 3 is defined as partially relevant, which means this grant is related to the users' research or they will try it in future. 1 means not relevant, which means the recommended grant is not useful for the researchers.

As shown in Table 6, with an average rate of 3.53 out of 5, the recommendation systems successfully provided relevant grant projects to the participants based on their input. And almost all of our participants found at least one project that interested them among the top 10 recommendations based on the highest rating score we received. These ratings prove our evaluation system to be efficient and helpful for researchers looking for relevant funded grant projects.

5. Discussion & conclusion

This work presented herein is the first of its kind toward developing a biomedical funded grant recommendation system for researchers

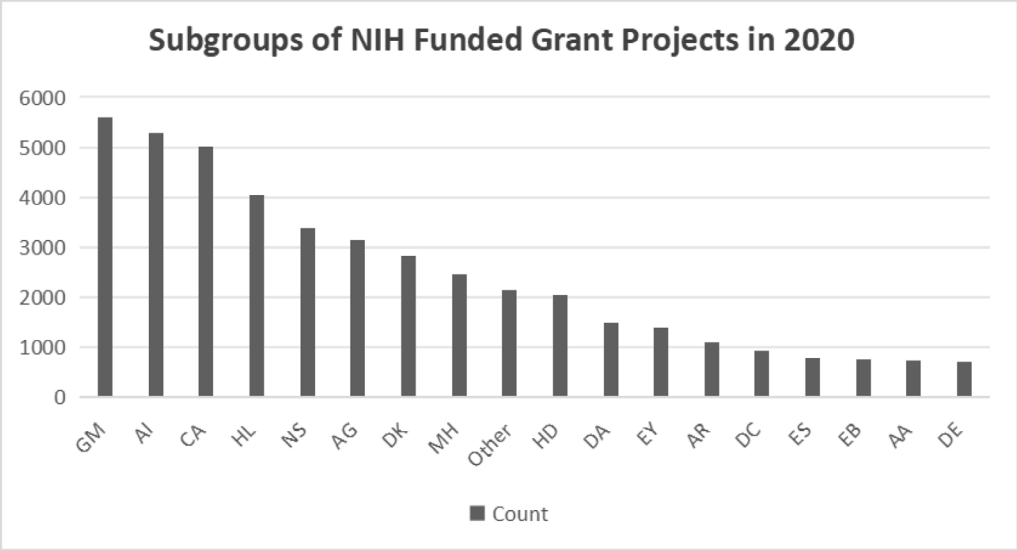


Fig. 3. Distribution of NIH Projects' Subgroups *. * Definition of Subgroup IDs and numeric statistics are listed in Appendix.

seeking grants and applications which may be related to their interested field of research. We lemmatized the textual data extracted from NIH funded grants related with research projects and their applications, and generated a grant recommender to compare the efficiency of several recommendation methods. Based on the internal evaluation results, pre-trained embedding model based on biomedical word and sentences significantly improve the precision of the recommendation. The external evaluation results from faculty members at our school also prove that such system is useful for biomedical researchers.

For the next step of this project, we aim to improve the flexibility of our recommendation, by developing a reasonable way to clustering the researchers' interests based on their publications and provide recommendation for each interest individually, which allowed us the flexibility to cluster each researcher's interests differently. We also plan to expand our NIH funded grants database, to generate public funded grant recommendation systems based on our system architecture, and to perform evaluation for our system by collecting further user feedback.

This research not only prove the efficiency of our proposed models compared to the traditional methods, but also highlights the importance of recommendation system in academia. We believe the requirement of making connections between researchers' interests and funded grant resources would grow rapidly with the development of biomedical research in the future.

CRediT authorship contribution statement

Zitong Zhang: Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ashraf Yaseen:** Writing – review & editing, Supervision, Conceptualization. **Hulin Wu:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The NIH funded grant project datasets are available as open data via the ExPORTER.

The FY 2020 RePORTER Project Data are available at https://reporter.nih.gov/exporter/projects.

The publication abstract data are available as open data via PubMed, at the website:

https://pubmed.ncbi.nlm.nih.gov/download/.

Appendix A. Abbreviation

Abbreviation	Explanation
BM	Best Matching
CBOW	Continuous Bag-of-words
GEO	Gene Expression Omnibus data
IR	Information Retrieval
NIH	the National Institute of Health
NSF	the National Science Foundation
TF-IDF	Term Frequency – Inverse Document Frequency
TP	True Positive
VRA	the Virtual Research Assistant platform
Abbreviation for Subgroups	
GM	General Medicine
AI	Allergy and Infectious Diseases
CA	Cancer
HL	Heart, Lung and Blood
NS	Neurological Disorders and Stroke
AG	Aging
DK	Diabetes, Digestive and Kidney Diseases
MH	Mental Health
HD	Child Health and Human Development
DA	Drug Abuse
EY	Eye
AR	Arthritis, Musculoskeletal and Skin Diseases
DC	Deafness and Communication Disorders
ES	Environmental Health Science
EB	Biomedical Imagine and Bioengineering
AA	Alcohol Abuse
DE	Dental

Appendix B. Distribution for NIH project subgroups

Filtering the 43,807 NIH research projects we extracted, 34 subgroups were obtained based on their research fields. The distribution of these subgroups is shown in the table below.

Subgroups	NIH Funded Grant Projects
General Medicine	5587
Allergy and Infectious Diseases	5275
Cancer	5020
Heart, Lung and Blood	4034
Neurological Disorders and Stroke	3390
Aging	3147
Diabetes, Digestive and Kidney Diseases	2823
Mental Health	2461
Other	2148
Child Health and Human Development	2046
Drug Abuse	1485
Eye	1394
Arthritis, Musculoskeletal and Skin Diseases	1108
Deafness and Communication Disorders	923
Environmental Health Science	787
Biomedical Imagine and Bioengineering	753
Alcohol Abuse	722
Dental	707
Total	43 810

References

Anon, 2014. SJFinder: SJFinder Recommend Journals. Retrieved from http://www.sjfinder.com/journals/recommend/.

Anon, 2023. RCDC: Categorization Process. National Institutes of Health, Retrieved from https://report.nih.gov/funding/categorical-spending/rcdc-process (Accessed on 2023).

Asabere, N.Y., Acakpovi, A., 2019. ROVETS: search based socially-aware recommendation of smart conference sessions. Int. J. Decis. Supp. Syst. Technol. 11 (3), 30–46.

Asabere, N.Y., Xu, B., Acakpovi, A., Deonauth, N., 2018. SARVE-2: exploiting social venue recommendation in the context of smart conferences. IEEE Trans. Emerg. Top. Comput. 9 (1), 342–353.

Benchettara, N., Kanawati, R., Rouveirol, C., 2010. A supervised machine learning link prediction approach for academic collaboration recommendation. In: Proceedings of the Fourth ACM Conference on Recommender Systems. pp. 253–256.

Bird, S., Klein, E., Loper, E., 2009. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. 5, 135–146.

Bollacker, K.D., Lawrence, S., Giles, C.L., 1998. CiteSeer: An autonomous web agent for automatic retrieval and identification of interesting publications. In: Proceedings of the Second International Conference on Autonomous Agents. pp. 116–123.

Chen, Q., Peng, Y., Lu, Z., 2019. BioSentVec: creating sentence embeddings for biomedical texts. In: In 2019 IEEE International Conference on Healthcare Informatics, ICHI. IEEE, pp. 1–5.

Errami, M., Wren, J.D., Hicks, J.M., Garner, H.R., 2007. eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications. Nucl. Acids Res. 35 (suppl_2), W12–W15.

ExPORTER, 2024. Retrieved from https://reporter.nih.gov/exporter.

Gollapalli, S.D., Mitra, P., Giles, C.L., 2012. Similar researcher search in academic environments. In: Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries. pp. 167–170.

Hong, K., Jeon, H., Jeon, C., 2012. UserProfile-based personalized research paper recommendation system. In: In 2012 8th International Conference on Computing and Networking Technology, INC, ICCIS and ICMIC. IEEE, pp. 134–138.

Kamada, S., Ichimura, T., Watanabe, T., 2015. Recommendation system of Grants-in-Aid for researchers by using JSPS keyword. In: 2015 IEEE 8th International Workshop on Computational Intelligence and Applications. IWCIA, IEEE, pp. 143–148.

Kamada, S., Ichimura, T., Watanabe, T., 2016. A recommendation system of grants to acquire external funds. In: 2016 IEEE 9th International Workshop on Computational Intelligence and Applications. IWCIA, IEEE, pp. 125–130.

Lapata, M., Blunsom, P., Koller, A., 2017. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol. 2, Short papers.

Lika, B., Kolomvatsos, K., Hadjiefthymiades, S., 2014. Facing the cold start problem in recommender systems. Expert Syst. Appl. 41 (4), 2065–2073.

MEDLINE, 2021. The National Library of Medicine's premier bibliographic database. Retrieved from https://www.nlm.nih.gov/medline/index.html/.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Adv. Neural Inform. Process. Syst. 26.

National Institution of Health, 2012. Bibliometric analyses of publications and grant productivity: SPIRES, a new Web-based tool. The National Institute of Environmental Health Sciences, The National Institutes of Health.

National Institution of Health, 2024. Institute/Center Code Definition for NIH projects.. Retrieved from https://grants.nih.gov/grants/glossary.htm.

NIH RePORTER, 2024. Retrieved from https://reporter.nih.gov/.

Patra, B.G., Maroufy, V., Soltanalizadeh, B., Deng, N., Zheng, W.J., Roberts, K., Wu, H., 2020a. A content-based literature recommendation system for datasets to improve data reusability–A case study on Gene Expression Omnibus (GEO) datasets. J. Biomed. Inform. 104, 103399.

Patra, B.G., Roberts, K., Wu, H., 2020c. A content-based dataset recommendation system for researchers—a case study on Gene Expression Omnibus (GEO) repository. Database baaa064.

Patra, B.G., Soltanalizadeh, B., Deng, N., Wu, L., Maroufy, V., Wu, C., et al., 2020b. An informatics research platform to make public gene expression time-course datasets reusable for more scientific discoveries. Database baaa074.

Perozzi, B., Al-Rfou, R., Skiena, S., 2014. Deepwalk: Online learning of social representations.. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 701–710.

PubMED, 2023. National Center of Biotechology Information, National Library of Medicine. Retrieved from https://pubmed.ncbi.nlm.nih.gov/about/.

Rajaraman, A., Ullman, J.D., 2011. Mining of Massive Datasets. Cambridge University Press.

Robertson, S., Zaragoza, H., 2009. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends® in Information Retrieval 3 (4), 333–389.

Tang, T.Y., McCalla, G., 2009. The pedagogical value of papers: a collaborative-filtering based paper recommender.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q., 2015. Line: Large-scale information network embedding. In: Proceedings of the 24th International Conference on World Wide Web. pp. 1067–1077.

U.S Department of Health and Human Services, 2023. National institutes of health, grants and funding. Achieved 2023, https://www.nih.gov/grants-funding.

Zhang, Y., Chen, Q., Yang, Z., Lin, H., Lu, Z., 2019. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci. Data 6 (1), 52.

Zhang, Z., Patra, B.G., Yaseen, A., Zhu, J., Sabharwal, R., Roberts, K., et al., 2023. Scholarly recommendation systems: a literature survey. Knowl. Inf. Syst. 1–46.

Zhang, Z., Yaseen, A., 2023. A content-based dataset recommendation system for biomedical datasets. In: 2023 International Conference on Information and Communication Technologies. ICICT, IEEE.

Zhu, J., Patra, B.G., Wu, H., Yaseen, A., 2023. A novel NIH research grant recommender using BERT. PLoS One 18 (1), e0278636.

Zhu, J., Patra, B.G., Yaseen, A., 2021. Recommender system of scholarly papers using public datasets. AMIA Summits Transl. Sci. Proc. 2021 (672).

Zhu, J., Wu, H., Yaseen, A., 2022. Sensitivity analysis of a BERT-based scholarly recommendation system. In: The International FLAIRS Conference Proceedings, vol. 35.

Zhu, J., Yaseen, A., 2022. A recommender for research collaborators using graph neural networks. Front. Artif. Intell. 5, 881704.