Bringing Artificial Intelligence to Research Analytics: Research Highlighter-MatchMaker

Tianning Huang
University at Albany, State University of New York

Yuemin Li
University at Albany, State University of New York

It is more significant to identify accurate grant opportunities given the increasing challenges in securing grants these days. Combining Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques, we have developed Research Highlighter-MatchMaker to facilitate the grant application process. The tool aims to provide an overview of research strength and to recommend matched funding opportunities as well as potential collaborators for researchers based on their research track records. It has the potential to expand to provide matching for on-campus researchers and industrial partners. It will also include a chatbot to facilitate the research administration infrastructure and process on campus.

Keywords: artificial intelligence, natural language processing, recommendation system, research enterprise, grant application

INTRODUCTION

Securing grants is essential for educational institutions, as it provides critical funding for research, fosters innovation, and supports educational programs. These grants empower academic communities to pursue new knowledge, offer students enhanced learning opportunities, and ultimately contribute to societal advancement and local community development. However, grant applications have become more and more competitive recently with the success rate has dropping significantly. For example, the National Institute of Health (NIH), the largest biomedical research agency in the world, has grants success rate dropping from approximately 30 percent in the late 1990s and early 2000s to 20 percent in 2023 (Belluz, et.al., 2016; National Institute of Health, 2024). The cut in grants budget by the federal government further exacerbates the competition. The National Science Foundation's budget this year was reduced by 8 percent, amounting to $6.6 billion (Palmer, 2024). Two high-profile NIH programs will also be cut by more than one-third, amounting to $462 million (Kaiser, 2024). Additionally, writing grants is rather time-consuming, with some estimates of taking up 50 percent of researchers' time (Lofgren, 2021). Another estimate presents an average proposal completion time for a single grant range from 170 hours to 270 hours, amounting to $8,500 to $13,400 in salary costs and $20,000 when including administrative overhead (Boyack, et.al., 2018). The decreased rate of successful grant applications, along with diminishing funding amounts and the escalated investment of time and effort required for the application process, have substantially escalated the

challenges in obtaining grants, making it increasingly vital to strategically dedicating time and resources to grant opportunities that align precisely with one's research expertise.

The big question is then how to identify the matched grant opportunities to increase success rate. A common practice by educational institutions is to provide funding opportunity databases, such as SPIN® or Pivot, together with other resources or directions to lists of grant opportunities for researchers. One issue with these listed resources is that their users, researchers, could be unaware of their existence and thus totally miss the benefits they could provide. Even when faculty and researchers are aware of the existence of these resources, it is overall very time consuming for them to go through the list one by one to identify the suitable grant opportunities. Additionally, most funding opportunity databases in the current market rely on keyword match algorithm in the backend to provide grant opportunities recommendations. Such mechanism could miss those relevant grant opportunities with no matched keywords, underscoring its effectiveness. Another caveat is that keyword matching algorithm does not take personal research track records into consideration while providing the recommendation, which could limit the recommendation accuracy. An even more challenging situation could exist when grant opportunity databases are fully unavailable due to resource limitations. Researchers will then need to rely on recommendations from colleagues and their own networks or search for suitable funding opportunities on their own. The downside of this scenario is that without a unified portal of funding opportunities, it is probable that researchers could miss good fundings simply because they are unaware of the existence of such opportunities. None of the above scenarios are ideal for researchers, considering the value of their time. They could be better off by saving those time spent on searching for suitable grant opportunities to more valuable research activities including writing the grant proposal or working on their research *per se*.

The Research Highlighter-MatchMaker project that we are proposing here could ideally address these challenges by providing a one-stop resource platform for researchers to find the matched grant opportunities without wasting their precious time on searching through a huge list nor missing any relevant funding opportunities simply because of unawareness. The project aims to build up research recommendation systems for researchers that recommend both targeted funding opportunities and potential multidisciplinary collaborations for researchers based on their publications and previous funding records using Artificial Intelligence (AI) and advanced Natural Language Processing (NLP) techniques. By considering researchers' research track records, the purpose of this project is to provide recommendations on more personalized and tailored grant opportunities as well as potential collaborators for researchers to save their time on searching and screening and to avoid the occurrence of missing good funding opportunities that they could have successfully obtained. The tool also provides an overview of research strengths of the whole institution for data-driven decision-making processes of the leadership. This project will have the potential to expand into recommendations on external collaborators including industrial partnerships with expansion of data sources such as technology transfer agreements or compliance documents. The project will also be expanded to include a chatbot function to facilitate the research administration process for higher education institutions with established infrastructure of research administration.

RESEARCH DESIGN

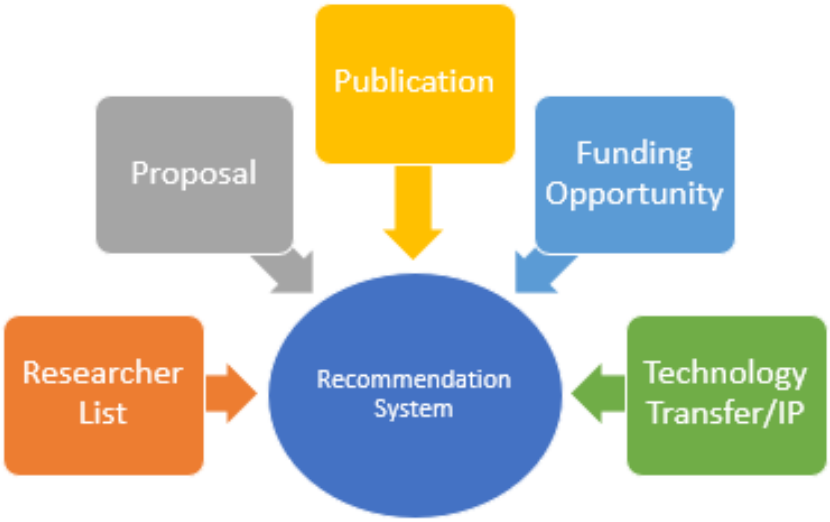
The core algorithm that lays the foundation of this project is an advanced NLP method called Topic Modeling, specifically BERTopic, which "leverages BERT embeddings and c-TF-IDF to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions" (Grootendorst, 2024). BERT, or Bidirectional Encoder Representations from Transformers, is a pre-trained open-sourced big language deep learning model introduced by researchers from Google in 2018 (Yang, et.al., 2019). One big advantage of BERTopic model is its capability of understanding text context. Previous topic modeling models such as the Latent Dirichlet allocation (LDA) model assign tokens into topics without the capability to take text context into consideration. BERTopic, with the BERT embedding function via sentence-transformers, can calculate semantic similarity for each document or sentence in the database fed into BERTopic (Egger & Yu 2022). This key step helps the model to identify sentences or documents with similar meanings before assigning them to topics, and thus has greatly improved the

outcome of the model. Another pros of BERTopic model is its flexibility to incorporate many Large Language Models (LLMs) including open-sourced Hugging Face algorithms, Meta Llama, and openAI/Chat-GPT (Grootendorst, 2024). One more benefit of BERTopic model lies in its open-source feature. Based on BERT, which is an open-source language model provided by Google, BERTopic is also open-source and does not charge fees for applications when we incorporate it to build the recommendation system.

Data

This project incorporates four types of data: (1) Researcher list, which contains identified researchers' biographical information including their name, email address, job title, department/unit, their campus ID, and Scopus ID (ORCID ID will be the next step); (2) Proposal abstracts and funding track records that are downloaded from the proposal database; (3) Publication abstracts that are downloaded from the Scopus, an abstract and citation database of peer-reviewed literature including scientific journals, books, and conference proceedings provided by Elsevier; (4) funding opportunities from SPIN®, a comprehensive database containing over 40,000 funding opportunities from more than 10,000 federal, public, non-profit, and private sponsors. A potential data source is the technology transfer/Intellectual Property (IP) data that will provide connections between academics, industry, and others.

FIGURE 1
DATA SOURCES

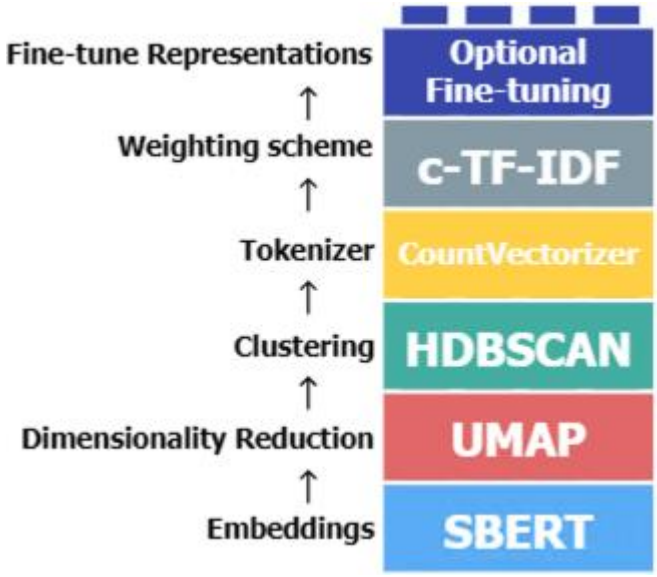


Data preprocessing for the BERTopic model includes data importing from different sources and preliminary data cleaning. In order to enable a more dynamic data importing process, we take advantage of APIs provided by SPIN® and Scopus and create our own APIs to import SPIN® data and Scopus data into the algorithm. Preliminary data cleaning includes removing empty records in the database, building filters on funding opportunities from SPIN® to include only eligible and current funding opportunities, and removing stop words such as "a", "the", and etc.

Algorithm

BERTopic model algorithm includes six layers, starting from embeddings, dimensionality reduction, clustering, tokenizer, weighting scheme and finally to fine-tune representations. Below we will discuss each layer in brief explanation and share fine-tuning parameters/algorithms that we include in our preliminary model. We tested each layer one by one and picked the parameter that gave us the most accurate results.

FIGURE 2
ALGORITHM



Embeddings

The first step in BERTopic converts text documents into numerical representations with the use of sentence-transformers. Sentence-transformers are a class of models and techniques used for generating embeddings (vector representations) for sentences or text passages. These embeddings capture the semantic meaning of the text and can be used for various Natural Language Processing (NLP) tasks. The default setting for BERTopic uses SBERT as the sentence-transformer (SBERT.net.). We tested different models and selected "allenai-specter", a Document-level Representation Learning using Citation-informed Transformers, for its high accuracy (Reimers & Gurevych, 2020; Allen Institute for AI).

Dimensionality Reduction

The second step is to reduce dimensionality of the converted text documents. Converting text documents into numerical vectors usually introduces high dimensionality, resulting in difficulties in clustering given the curse of dimensionality. BERTopic uses UMAP (Uniform Manifold Approximation and Projection) as a default solution to reduce the dimensionality of the embeddings to a workable dimensional space for clustering algorithms to work with. UMAP can capture both the local and global high-dimensional space in lower dimensions while preserving the underlying structure and relationships between data points. We tested both UMAP and another dimension reduction method, PCA (Principal Component Analysis), deciding that UMAP produced better clustering results. UMAP is superior for two reasons: 1) UMAP aims to preserve both local and global structure within the data. Local structure refers to the relationships between nearby data points, while global structure refers to the broader patterns and clusters in the data; 2) Unlike PCA, UMAP is nonlinear, meaning it can capture complex, nonlinear relationships in the data, therefore it is suitable for a wide range of datasets, including those with intricate geometric structures.

Clustering

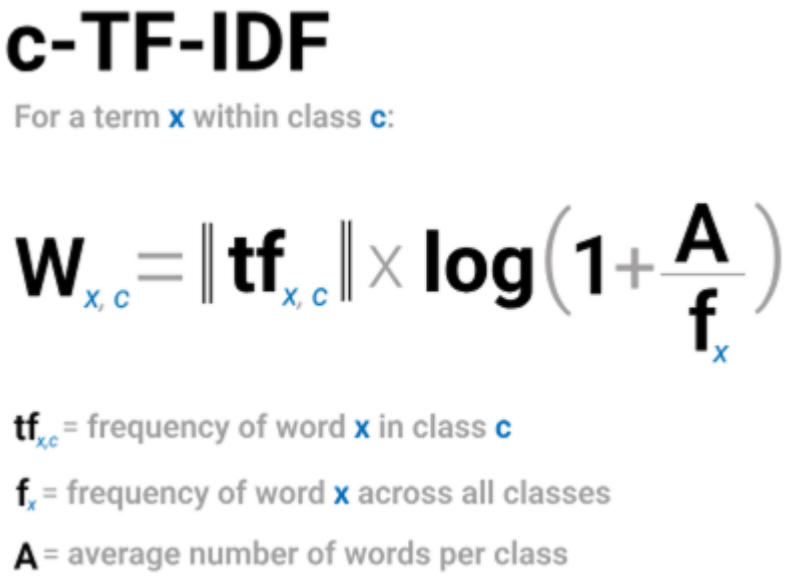
The third step is to cluster the dimension reduced input embeddings into groups of similar embeddings to extract the topics. We compared two clustering methods: HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) and K-Means. HDBCAN is a density-based clustering algorithm, meaning it identifies clusters by grouping together data points that are close to each other in a high-density region while considering outliers as noise points. Instead, K-Means is a centroid-based

clustering algorithm that it partitions data points into clusters by minimizing the sum of squared distances between data points and the cluster centers (centroids). HDBSCAN returns a better performance regarding to the clustering of research text data, therefore, we select HDBSCAN as the clustering algorithm in our BERTopic model.

Tokenizer and Weighting Scheme

The next step is called Tokenizer, which creates the topic representations. BERTopic model uses CountVectorizer, a text preprocessing technique to convert a collection of text documents into a matrix of token (word) counts. We set the number of tokens (words) for each entity in a topic representation in a range of one to three. We include both a default and a self-defined stopwords dictionary. The self-defined stopwords dictionary is built up based on input text data features. Next, BERTopic uses c-TF-IDF to get an accurate representation of the topics from our bag-of-words matrix. This measurement takes into consideration what makes the documents in one cluster different from documents in another cluster.

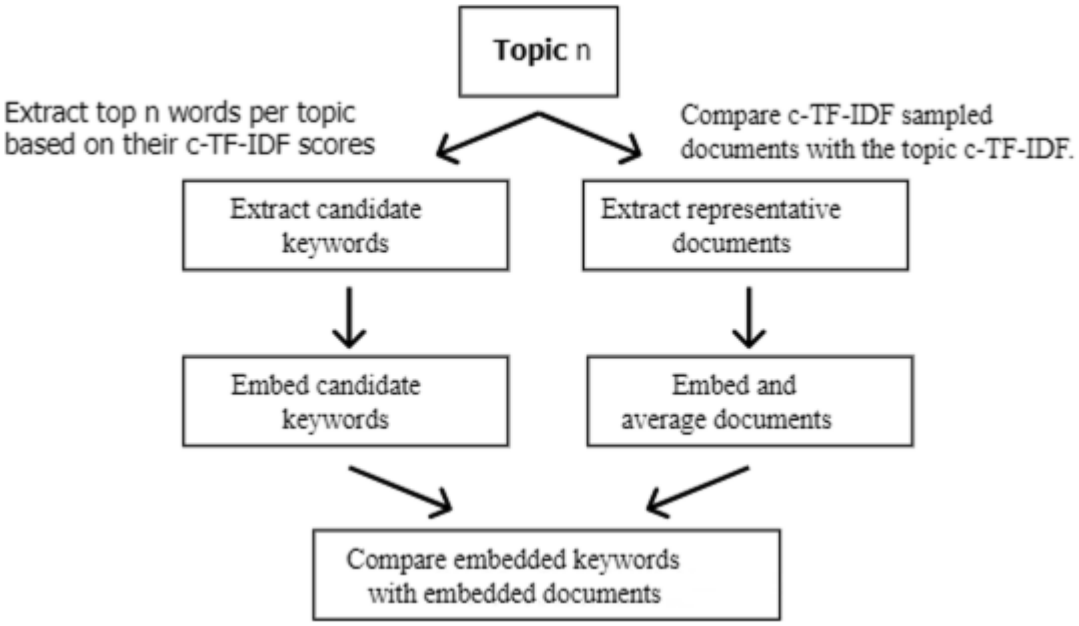
FIGURE 3
C-TF-IDF



Fine-tune Representations

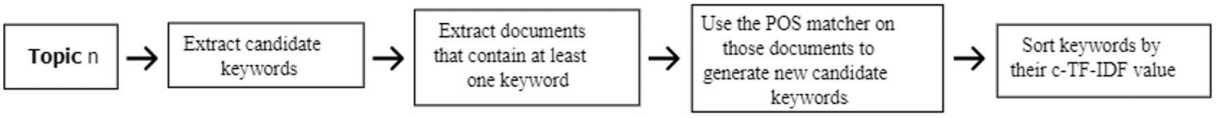
The last step is to fine-tune the topic representation with the help of representation models. The purpose of this step is to obtain the representative topics based on the semantic relationship between keywords/keyphrases and the set of documents in each topic after generating the topics with c-TF-IDF. A KeyBERTInspired representation model is used to first leverage c-TF-IDF to create a set of representative documents per topic. These documents are used as the updated topic embedding to calculate the similarity between candidate keywords and the topic embedding.

FIGURE 4
TOPIC REPRESENTATION



We also add two more parameters in the representation model to improve its accuracy after testing: PartOfSpeech and MaximalMarginalRelevance. By introducing PartOfSpeech, we identify only noun keywords and exclude keywords that are verbs or adjectives from the representation keywords for the topics generated by the model. Limiting to noun keywords will help identify keywords that are most relevant to identify the topic of specific research fields recognized by the model.

FIGURE 5
TOPIC REPRESENTATION - POS



Additionally, an algorithm called Maximal Marginal Relevance (MMR) is adopted to decrease the redundancy and improve the diversity of representation keywords. The algorithm considers the similarity of keywords/keyphrases with the document, along with the similarity of already selected keywords and keyphrases. This results in a selection of keywords that maximize their within diversity with respect to the document.

FIGURE 6
TOPIC REPRESENTATION - MMR

$$MMR = arg \max_{D_i \in R \setminus S} [\lambda Sim_1(D_i, Q) - (1 - \lambda) \max_{D_j \in S} Sim_2(D_i, D_j)]$$

Algorithm Improvement

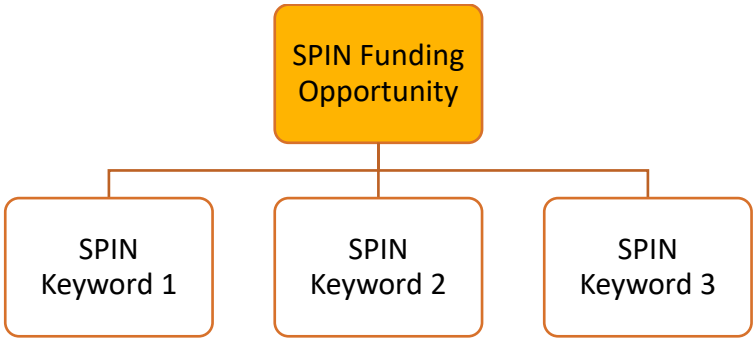
After going through the above steps, BERTopic method calculates probability score of each research document under each topic cluster and assigns that research document to a topic cluster with the highest

probability score. Although this default step returns good results for general topic modeling tasks, it introduces a caveat into our recommendation system. Assigning one research document to only one topic according is problematic because a researcher can have more than one research interest. When a researcher has fewer research documents in the database, assigning her documents to only one topic could lead to underestimation of the width of a researcher's research interests. In order to address this gap in the method, we adopt a non-traditional BERTopic step to sum up probability score of each topic cluster for each faculty/researcher in the database and identify topic clusters that is greater than a threshold summed probability score. With this improvement, we are able to assign one researcher to multiple research topics.

Another modification that we adopt is to add a second layer of topic modeling for large topic clusters that contain the number of research documents greater than a threshold. The purpose of this modification is to address the issue of granularity. It is the feature of the clustering algorithm HDBSCAN that it tends to return large topic clusters. Such a feature could lead to getting too broad research area when it needs to be more granular. By running BERTopic on large topic clusters as a second layer, the modified model would further break down these clusters into smaller clusters and thus provide a more granular result.

A third modification is keyword match recommendation. Besides BERTopic modeling, we plan to add a keyword match algorithm based on SPIN® keywords. SPIN® has a hierarchy system of keywords and each of its funding opportunity system is tagged with a set of keywords. Our algorithm first searches through the input research documents to identify matched SPIN® keywords in the research documents. The algorithm then calculates a matching score for each matched SPIN® keyword by multiplying the number of times the match occurs, the priority of author's position (if she is first author, multiply 3), and the ratio of matched documents out of total input research documents. For example, as for a SPIN® keyword "AI", it appears 3 times in a researcher's 1 out of 40 records as the first author, the matching score will be (100+50+25) x 3 x 1/40 = 13.125. The algorithm will then recommend funding opportunities based on keywords with high matching scores.

FIGURE 7
KEYWORD HIERARCHY SYSTEM



A fourth improvement that we will adopt is adding a text search algorithm based on BERTopic model results. When a researcher or a student has a random text data, she will be able to insert this text data into the recommendation system to obtain recommendations on funding opportunities and collaborations within research clusters. To achieve this outcome, the text search algorithm will be based on BERTopic model outputs, and it will assign any newly input text data to topics that are identified by the trained model. The algorithm will then report the topic area that the newly imported text data belongs to, the faculty/researcher who shares similar research interests, and the funding opportunities that are matched with the searched text.

Expected Outcomes/Deliverables

A major deliverable will be a front-end portal of the funding opportunity recommendation system that consists of three major buttons: "Search By Name", "Search By Topic", and "Search By Text". Within the "Search by Name" page, users will be able to search faculty/researcher names to obtain their research area

and funding opportunities recommended by the recommendation system. Within the "Search By Topic" page, users can type and search a specific topic area and obtain a list of researchers that are grouped under that topic by the recommendation system. The list contains faculty/researcher name, email address, school, department, and their research documents that are identified by the recommendation system under that specific topic. This function provides recommendations on potential collaborations for faculty/researcher who are grouped into one cluster by the algorithm. Under the "Search By Text" button, users will be able to input any texts and the recommendation system will recommend funding opportunities based on input texts. With this function, any research text/document that is not associated with faculty/researcher in the database could be fed into the recommendation system to obtain funding opportunity recommendations, which could greatly broaden the users of the recommendation system. The recommendation system will not only benefit faculty/researchers whose research records are included in the database, but also help other researchers, i.e. graduate students, to connect to research opportunities. Associated with the front-end portal is a regular email recommendation listserv which will send out relevant funding opportunities recommendations through emails to faculty/researchers.

Another deliverable will be a faculty/researcher network. The network will present connections between faculty/researchers based on their research documents and clusters generated by the recommendation system. Such a network could not only depict connections of faculty/researchers assigned to the same topic cluster by the recommendation system, but also suggest current or potential collaborations for faculty/researchers appearing in the same cluster. Given the feature of R1 institutions (Carnegie Classification of Institutions of Higher Education), the recommendation system will include more abundant research data related to faculty/researchers within the R1 institutions, but fewer research records for faculty/researchers residing in less research-intensive institutions. To address this gap, the faculty/researcher network will not only show information on connections among shared research interests but also include funding records of faculty/researcher. Therefore, when faculty from less research-intensive institutions search for potential collaborations with R1 institutions, they will be able to identify faculty/researchers with shared research interests and to collaborate with faculty/researchers who have experiences in applying or obtaining a specific funding that they might have in mind.

Potential Development

One potential development of the Research Highlighter-MatchMaker project is to provide a chatbot toolbox to guide Research Administration processes on campus for higher education institutions. First, such a chatbot could help answer standardized questions related to Research Administration processes ranging from proposal submission to travel reimbursement on awards based on the current set-up policy and regulations. Second, when the chatbot figures out that the question is out of its capability or people would like to communicate directly with Research or Grant Administrators, they could also use the chatbot to turn to the correct person of contact. One big challenge in research infrastructure is that faculty and researchers usually find it challenging to connect to the staff who have the answer to their questions, which could be time consuming for both faculty/researchers and staff. A chatbot leveraged on LLMs and AI could help address this issue and save time and efforts in this communication challenge. A third function of the chatbot is its assistance in paperwork preparation for Research Administration, from preparing a checklist of required documents of a proposal to guiding people to fill out travel reimbursement form, which will further facilitate the ease of communication and reduce the burdens for both faculty/researchers and staff.

Another potential expansion is to use the algorithms to connect scientific innovations and discoveries on campus, such as technology transfer, patents, and Intellectual Property (IPs) with industry partners or investment firms. Similar to provide funding opportunity recommendations to researchers, the project could offer recommendation matching for industry partners and investment firms such as venture capitalists if they are interested in cooperating to commercialize a certain innovation in technology or product developed by researchers on campus. They could also use the search buttons to search for researchers who develop researches they are interested in or for research fields that they would like to invest to commercialize.

DISCUSSION AND CONCLUSION

Overall, the recommendation system in the package will greatly enhance the infrastructure for research and education, not only at University at Albany campus per se, but has the potential to develop and spread into higher education institutions nationally. Given its automatic feature, users of the recommendation system will not be required to possess knowledges in NLP/ML techniques. Instead, they will be able to benefit from this recommendation system by simply typing their searching keyword/texts. The administrator and the leadership of higher education institutions will be able to obtain an overview of research clusters within the institution, including which research areas are core research interests of the institution, which faculty/researcher are included in different research clusters, as well as collaboration/connection networks among faculty/researcher. This project will benefit both research-intensive colleges such as Carnegie R1 institutions, but especially for institutions with fewer resources for its ease of use and low use barriers.

This system will also help increase partnerships between academia, industry and others. The faculty/researcher could take advantage of this system by identifying tailored funding opportunities without spending a large amount of time searching around different database or filtering relevant fundings one by one on their own. They will also be able to find potential collaborations with the help of the recommendation system and the associated network. Students will also benefit from the recommendation system to search for faculty/researcher they would like to connect and find relevant funding opportunities that are eligible for them. Industry collaborators, current or potential, could use the recommendation system to search for faculty/researcher for potential collaborations and identify potential technology transfer opportunities.

REFERENCES

Allen Institute for AI. (n.d.). Specter. GitHub repository. Retrieved from https://github.com/allenai/specter

Belluz, J., Plumer, B., & Resnick, B. (2016, September 7*). The 7 biggest problems facing science, according to 270 scientists.* Retrieved from https://www.vox.com/2016/7/14/12016710/science-challeges-research-funding-peer-review-process#1

Boyack, K., Smith, C., & Klavans, R. (2018). Toward Predicting Research Proposal Success. *Scientometrics*. DOI: 10.1007/s11192-017-2609-2.

Carnegie Classification of Institutions of Higher Education. (n.d.). Retrieved from https://carnegieclassifications.acenet.edu/

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, *7*(886498). https://doi.org/10.3389/fsoc.2022.886498

Grootendorst, M. (2024). *BERTopic Documentation*. Retrieved from https://maartengr.github.io/BERTopic/api/bertopic.html

Lofgren, E. (2021, December 28). *Top researchers spend 50% of their time writing grants: How to fix it and what it means for DoD.* Retrieved from https://acquisitiontalk.com/2021/12/top-researchers-spend-50-of-their-time-writing-grants-how-to-fix-it-and-what-it-means-for-dod/

Kaiser, J. (2024, April 30). *Major budget cuts to two high-profile NIH efforts leave researchers reeling.* Retrieved from https://www.science.org/content/article/major-budget-cuts-two-high-profile-nih-programs-leave-researchers-reeling

National Institute of Health. (2024). *Success Rates: R01-Equivalent and Research Project Grants.* Retrieved from https://report.nih.gov/nihdatabook/category/10

Palmer, K. (2024, May 2). *Researchers 'Shocked and Disappointed' After NSF Budget Cuts.* Retrieved from https://www.insidehighered.com/news/government/science-research-policy/2024/05/02/researchers-shocked-and-disappointed-after-nsf

Reimers, N., & Gurevych, I. (2020). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *arXiv*. preprint arXiv:2004.07180.

SBERT.net. (n.d.). *Pretrained models*. Retrieved from https://www.sbert.net/docs/sentence_transformer/pretrained_models.html

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q.V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv*. preprint arXiv:1810.04805.