Predicting Research Funding Success from Grant Proposal Abstracts

ECO482

Unurjargal Batsuuri / unu.batsuuri@mail.utoronto.ca / 1007766703

Joyce Lin / joycetheok.lin@mail.utoronto.ca / 1007957285

Daniella Chung / daniella.chung@mail.utoronto.ca / 1007687120

INTRODUCTION

MOTIVATION

Research is the lifeblood of progress. It creates innovations that save lives—think of insulin, or penicillin—and tackles the most pressing issues we face today. For instance, research on depression has significantly improved the effectiveness of the diagnosis and treatment of depression in patients (Trivedi, 2020), and research on the LGBTQ+ community has shed light on the stigma and hostility that those in that community faced (Koch et al., 2023). It drives global economies, creates new industries, and influences policy that improves welfare for all.

However, traditional grant proposal evaluation is time-consuming, subjective, and unable to fully account for the societal and economic impact of the proposed research. As a result, high-potential projects may go underfunded, delaying progress. The cost of inaction is incalculable. Thus, we aim to leverage machine learning and textual analysis to identify the key features that predict the likelihood of a grant proposal abstract receiving higher funding.

RESEARCH QUESTION

Our research question is: *Can the textual features of grant proposal abstracts predict funding outcomes in Canadian health research?* Although machine learning is quite new and requires constant validation and training to ensure that its analysis remains as objective as possible (Ueda et al., 2021), it has the potential to significantly reduce time spent reviewing and reduce human bias in the evaluation process. Thus, our research can make the grant review process more objective, efficient, and transparent.

LITERATURE REVIEW

Our literature seeks to highlight the current state of textual analysis usage in fund allocation and identify gaps in the literature that our research aims to fulfill.

In “On Predicting Research Grants’ Productivity Via Machine Learning,” Tohalino and Amancio (2022) employ machine learning to predict grant productivity for grants in medicine, dentistry, and veterinary medicine and find that the topic and year of publication are the most significant predictors. Whereas the paper analyzes productivity using bibliometric statistics, we aim to analyze the amount of funding received using the grant abstract.

Lupyani and Phiri’s (2024) paper “From Algorithms to Grants: Leveraging Machine Learning for Research and Innovation Fund Allocation” explores the application of machine learning to improve the grant allocation process. Although this paper aims to automate the evaluation of proposals as well, it implements text classification algorithms instead of regression.

In “A Data-Driven Approach in Predicting Scholarship Grants of a Local Government Unit in the Philippines Using Machine Learning,” Fajardo et al. (2024) develops a machine learning model to match scholarship applicants in the Philippines with the best scholarship, and finds logistic regression to be the best-performing model. The goal of predicting fund (scholarship) application is similar to ours, but we primarily predict with the textual characteristics of the grant, while the paper uses the applicants’ characteristics and background.

DATA AND METHODS

DATA

We scraped the dataset used in the research on November 17th, 2024 from the Canadian Institutes of Health Research (CIHR) Funding Decision Database using Selenium and BeautifulSoup. The dataset contains 7,011 proposals within the Social, Cultural, Environmental, and Population Health theme across the USA and Canada, with the following key variables: the institutions responsible for the project, project abstracts, and the funding amount provided by CIHR (CIHR contribution), which is our outcome variable. The dataset contains no missing data, although approximately 1% of the dataset was identified as duplicates during initial data analysis.To reduce bias and increase prediction, 30 outliers were removed from the dataset. Additionally, the CIHR Contribution variable had a highly skewed distribution and was log-transformed to achieve a more normal distribution for the analysis.

METHODOLOGY

Text preprocessing

Text preprocessing was conducted on the collected abstracts to clean and prepare the text for analysis, which involved tokenizing the text, removing irrelevant characters, lemmatizing, and extracting bigrams using the Gensim library. This process ensured that the text data was structured and meaningful for further analysis.

Topic Modeling

Latent Dirichlet Allocation (LDA) was employed to extract latent themes from the project abstracts. The optimal number of topics was determined to be 5 based on the coherence scores, which measure the semantic similarity of words within each topic. The topic modeling process produced topic distributions for each abstract, which were subsequently added in the dataset as additional features for predictive modeling.

Feature Engineering

In addition to the topic distributions, other features were engineered to enrich the dataset. Novelty scores were calculated to quantify the uniqueness of each abstract in comparison to others, using cosine similarity. The novelty scores ranged from 0 to 1, providing a measure of textual uniqueness. Furthermore, the duration of project terms, length of the abstract were incorporated to improve the predictive power of the models.

Train-Test Split and Data Scaling

The dataset was split into training and testing subsets, with 80% of the data used for training and 20% reserved for testing. This split ensured and unbiased evaluation of model performance on unseen data. Numerical features were scaled using the StandardScaler function, which normalized the data and improved the performance of the models during training and evaluation.

Model Training

A variety of machine learning algorithms were tested to identify the best performing model for predicting CIHR contributions. These included linear models such as Linear Regression, Lasso, Ridge, and Elastic Net; tree based models such as Decision Tree, Random Forest, and Gradient-Boosting trees; and K-Nearest Neighbors (KNN). To optimize performance, the hyperparameters of the models were tuned using GridSearchCV with five-fold cross-validation, ensuring the models were fine-tuned to their best configuration.

Model Evaluation

The trained models were evaluated using Mean Squared Error (MSE) to measure prediction error and R^2 score to assess the proportion of variance explained by the model. These metrics provided insights into the accuracy and reliability of each model's predictions, allowing for the selection of the most effective approach for predicting funding amounts.

RESULTS

Topic Modeling Results

Latent Dirichlet Allocation (LDA) topic modelling identified five distinct themes within the project abstracts that can be seen in Figure 1, with an optimal coherence score of 0.325, indicating meaningful differentiation among topics. The themes were categorized as follows:

1. *Topic 1 - **Women's Health in Canada***

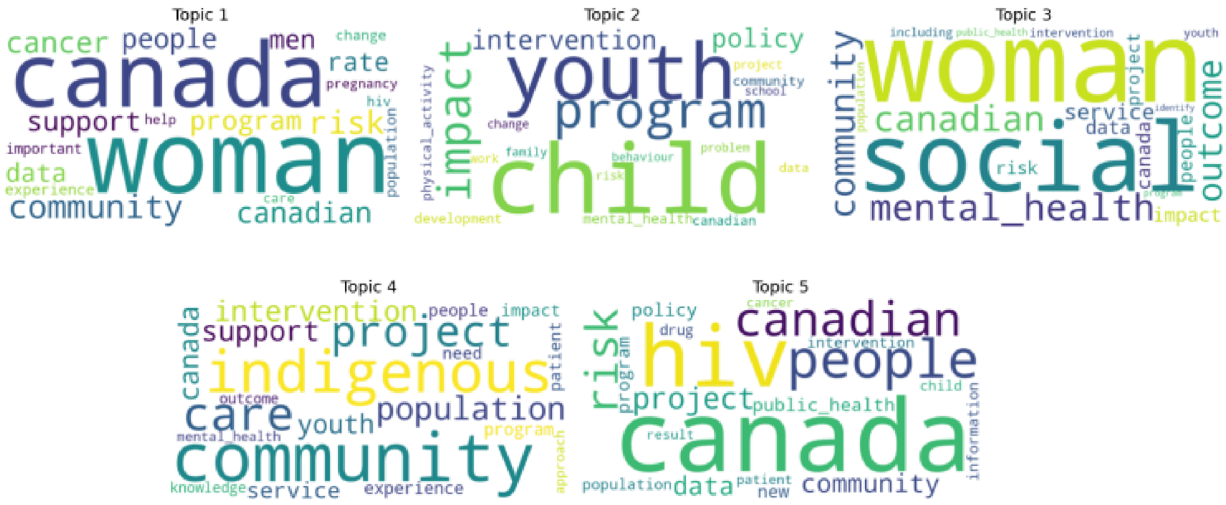
2. *Topic 2 - **Youth-Centered Policy***

3. *Topic 3 - **Mental Health and Social Outcomes***

4. *Topic 4 - **Indigenous Community Health***

5. *Topic 5 - **HIV Risk and Canadian Population Health***

Figure 1. WordCloud of the 5 topics



Each project in the dataset was assigned a topic distribution, representing its relevance to each of the five themes. These distributions served as critical features for understanding funding patterns and predicting grant allocations.

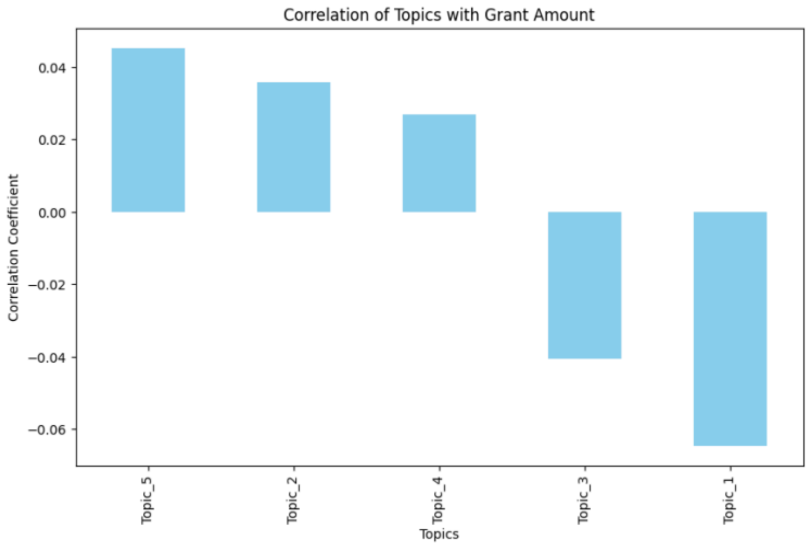
Model Evaluation Results

The Gradient Boosting model was the best-performing predictive model, achieving an R^2 score of **0.5714** and an MSE of **1.4496**, outperforming all other models. Random Forest and Decision Tree models were the next best performing models, with R^2 scores of 0.5590 and 0.5241, respectively. Linear models, except for Linear Regression, performed similarly, with R^2 scores of approximately 0.45 and MSE values around 1.85. On the other hand, Linear Regression produced a negative R^2 value, indicating that the data does not exhibit a strong linear relationship. Lastly, K-Nearest Neighbors (KNN) underperformed, with an R^2 score of 0.3312, revealing limited predictive power in capturing the complex relationships present in the data.

In summary, while the Gradient Boosting model demonstrated the best overall performance, its predictive power, with an R^2 of 0.5714, remains limited. By combining the benefits of decision trees with boosting techniques, Gradient Boosting effectively captured the intricate dependencies within the dataset, making it the most robust option for predicting grant amounts. However, despite its strong performance relative to other models, the R^2 value of 0.5714 indicates that the model explains only 57.14% of the variance in grant amounts, leaving a considerable portion (42.86%) of the variability unexplained. This suggests that grant allocation decisions may depend on additional unobserved factors or interactions not captured by the current set of features.

Correlation Between Topics and Grant Amounts

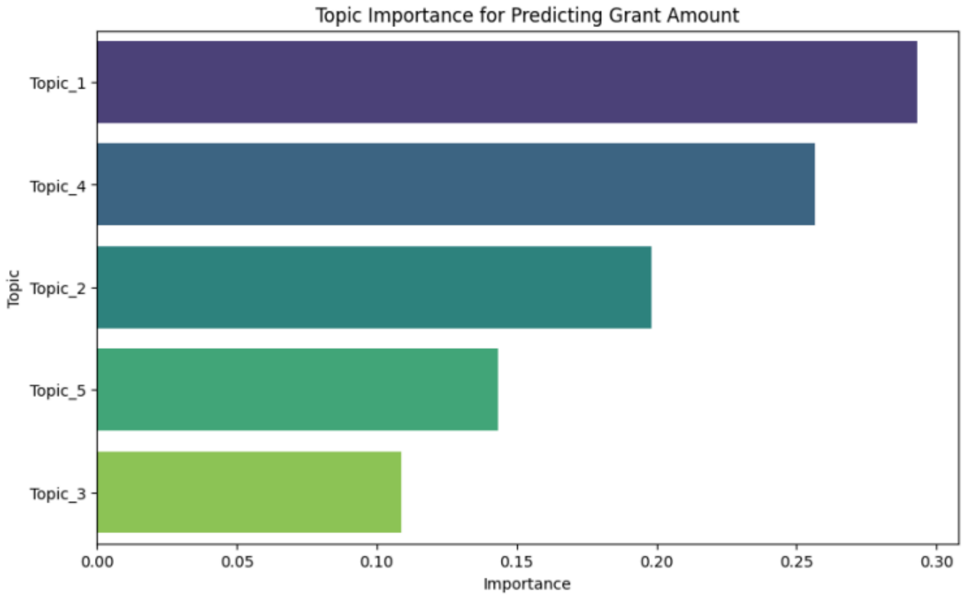
Figure 2. Correlation of Topics with Grant Amount



The relationship between topics and grant amounts, visualized in Figure 2, revealed notable variations in funding priorities. Topic 5 (HIV Risk and Canadian Population Health) exhibited the strongest positive correlation with grant amounts, highlighting significant support for public health initiatives addressing HIV risk. Similarly, Topic 4 (Indigenous Community Health) showed a positive correlation, reflecting a prioritization of Indigenous health issues in funding decisions. Topic 2 (Youth-Centered Policy) displayed a moderate positive correlation, suggesting steady support for youth-related projects. In contrast, Topic 1 (Women's Health in Canada) demonstrated the strongest negative correlation with grant amounts, suggesting systemic underfunding of women's health initiatives. Topic 3 (Mental Health and Social Outcomes) also exhibited a negative correlation, though weaker, indicating lower financial support for mental health projects. These findings highlight the variability in funding priorities, with certain themes, such as HIV risk and Indigenous health, receiving stronger financial backing, while others, such as women's health and mental health, face greater funding challenges.

Predictive Power of Topics

Figure 3. Topic Importance for Predicting Grant Amount



The feature importance analysis from the Gradient Boosting model (Figure 3) highlighted the critical role of topic distributions in predicting grant amounts. Topic 1 (Women's Health in Canada) emerged as the most significant feature, despite its negative correlation with funding, emphasizing its distinct patterns and influence on model predictions. Topic 4 (Indigenous Community Health) ranked as the second most important predictor, aligning with its positive correlation and reflecting its prioritization in funding decisions. Topic 2 (Youth-Centered Policy) and Topic 5 (HIV Risk and Canadian Population Health) also contributed significantly to the model. Topic 5, in particular, reinforced its importance due to its strong positive correlation with higher grant amounts. In contrast, Topic 3 (Mental Health and Social Outcomes) exhibited the least predictive power, consistent with its weaker correlation, suggesting that mental health projects are less prioritized in funding decisions. These findings highlight how topic modeling, when integrated into machine learning frameworks, can reveal meaningful patterns and offer valuable insights into grant allocation dynamics. The predictive contributions of topics like Indigenous health and HIV risk reflect current funding priorities, while the weaker impact of mental health highlights areas that may require greater attention.

Key Findings

The results revealed significant insights into funding allocation patterns and predictive modeling performance. Topic modeling identified five distinct themes, with HIV Risk and Canadian Population Health and Indigenous Community Health exhibiting positive correlations with grant amounts, reflecting prioritization of public health and Indigenous health initiatives. Conversely, Women's Health in Canada and Mental Health and Social Outcomes showed negative correlations, highlighting systemic underfunding of these areas. These disparities suggest opportunities to reevaluate funding strategies to address unmet needs. Furthermore, Gradient Boosting was the best-performing predictive model, achieving an R^2 of 0.5714 and an MSE of 1.4496, outperforming other models such as Random Forest and Decision Tree. Feature importance analysis highlighted the critical role of topic distributions, with Women's Health in Canada being the most influential predictor despite its negative correlation with funding. While the model effectively captured key patterns, its limited R^2 score indicates that other unobserved factors likely influence grant allocation decisions. These findings underscore the potential of combining topic modeling and machine learning to uncover meaningful insights into funding priorities.

CONCLUSION

It is imperative that healthcare grant proposals receive efficient funding, as these research projects further the frontier of scientific research and work to better the standard of living for all. Our project revealed systematic underfunding in Women's Health and Mental Health fields, with the most funding being allocated to Indigenous Health. This reveals a concerning potential bias against marginalized groups such as women and those suffering from mental health issues, which may prevent these fields from developing life altering advancements in care.

LIMITATIONS

There were many limitations on our project, the first being the scope of our research. We used data from the Social, Cultural, Environmental, and Population Health theme from the CIHR database, leaving other themes in Health Research such as Biomedical or Clinical research unexplored. Further, we ran into monetary constraints. When exploring the use of a sentiment analysis, our first choice of lexicon was the linguistic inquiry and word count (LIWC), as it is suited for academic content. However, it was locked behind a paywall and we were unable to use it. Additionally, the makeup of the grant committee reviewing the proposals is likely to impact the amount of funding allocated, but there was no data available on these individuals. Taking their demographics and characteristics into account could explain some of the 43% of variation left unexplained by our gradient boosting model. This lack of available data extends to grant proposals that were considered and did not receive any funding - it is common practice for these applications not to be publicly listed, and as such, there is missing data on this class of grant proposal.

FURTHER RESEARCH

There are multiple avenues to explore in order to uncover more information regarding grant allocation, the first of which involves topic relevance. If a proposed research topic is related to a hot topic in public opinion or a popular field in scientific communities, would it increase the amount of funding the grant receives? Further, the applicant's qualifications and demographics could play an important role in grant determination. It follows that the grant committee would take notice of these factors, as a history of success indicates future success and may be awarded with more funding, and investigation may reveal bias based on gender or nationality. Finally, it would be prudent to follow up and observe the productivity of grants awarded in order to assess if the money is being efficiently allocated. If more funding is allocated to projects that have substandard performance, the allocation process should be refined to correct this inefficiency. Our project serves as a glimpse into the inner workings of grant allocations and the initiation of important healthcare research. Research investigating the questions listed above will reveal more information and may lead to more specific and actionable conclusions regarding the allocation of funds based on grant proposals.

CITATIONS

- A. Fajardo, R. C., Yara, F. B., Ardeña, R. F., Hernandez, M. K., & T. Arroyo, J. C. (2024). A data-driven approach in predicting scholarship grants of a Local Government Unit in the Philippines using Machine Learning. *International Journal of Engineering Trends and Technology*, 72(6), 74–81. https://doi.org/10.14445/22315381/ijett-v72i6p108
- Government of Canada, C. I. of H. R. (2018, January 24). *Funding decisions database*. CIHR. https://webapps.cihr-irsc.gc.ca/decisions/p/main.html?lang=en#fq={!tag=theme2}theme2%3A%22Social%20%2F%20Cultural%20%2F%20Environmental%20%2F%20Population%20Health%22&fq={!tag=country}country%3ACanada%20%20%20OR%20%20%20country%3A%22United%20States%20of%20America%22&sort=namesort%20asc&start=0&rows=20
- Koch, A., Rabins, M., Messina, J., & Brennan-Cook, J. (2023). Exploring the challenges of sexual orientation disclosure among lesbian, gay, bisexual, transgender, Queer Individuals. *The Journal for Nurse Practitioners*, 19(10), 104765. https://doi.org/10.1016/j.nurpra.2023.104765
- Lupyani, R., & Phiri, J. (2024). From algorithms to grants: Leveraging Machine Learning for Research and Innovation Fund Allocation. *Lecture Notes in Networks and Systems*, 469–480. https://doi.org/10.1007/978-3-031-54820-8_38
- Tohalino, J. A. V., & Amancio, D. R. (2022). On predicting research grants productivity via machine learning. *Journal of Informetrics*, 16(2), 101260. https://doi.org/10.1016/j.joi.2022.101260
- Trivedi, M. (2020, December 9). *Breakthroughs in depression research lead to more effective treatment*. UT Southwestern Medical Center. https://utswmed.org/medblog/antidepressants-research-treatments/
- Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., Matsui, Y., Nozaki, T., Nakaura, T., Fujima, N., Tatsugami, F., Yanagawa, M., Hirata, K., Yamada, A., Tsuboyama, T., Kawamura, M., Fujioka, T., & Naganawa, S. (2023). Fairness of Artificial Intelligence in Healthcare: Review and recommendations. *Japanese Journal of Radiology*, 42(1), 3–15. https://doi.org/10.1007/s11604-023-01474-3