

Predicting Research Funding Success from Grant Proposal Abstracts

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

Joyce Lin / joycetheok.lin@mail.utoronto.ca / 1007957285 / linjin42

Daniella Chung / daniella.chung@mail.utoronto.ca / 1007687120 / chungd25

GROUP 3

Motivation



On the Social Significance of Research

Research is the lifeblood of **progress**. It creates innovations that save lives and tackles the most pressing issues we face today. It can drive global economies, create new industries, and influence policy that improves welfare for all.

The Challenge:

Traditional grant proposal evaluation is

- **time-consuming**
- **subjective**
- **unable to fully account** for the societal and economic **impact** of proposed research

As a result, high-potential projects may go underfunded, delaying progress.

The **cost** of **inaction** is **incalculable**.



Our Approach & Contribution

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.

Machine learning has the potential to: (Ueda et al., 2021)

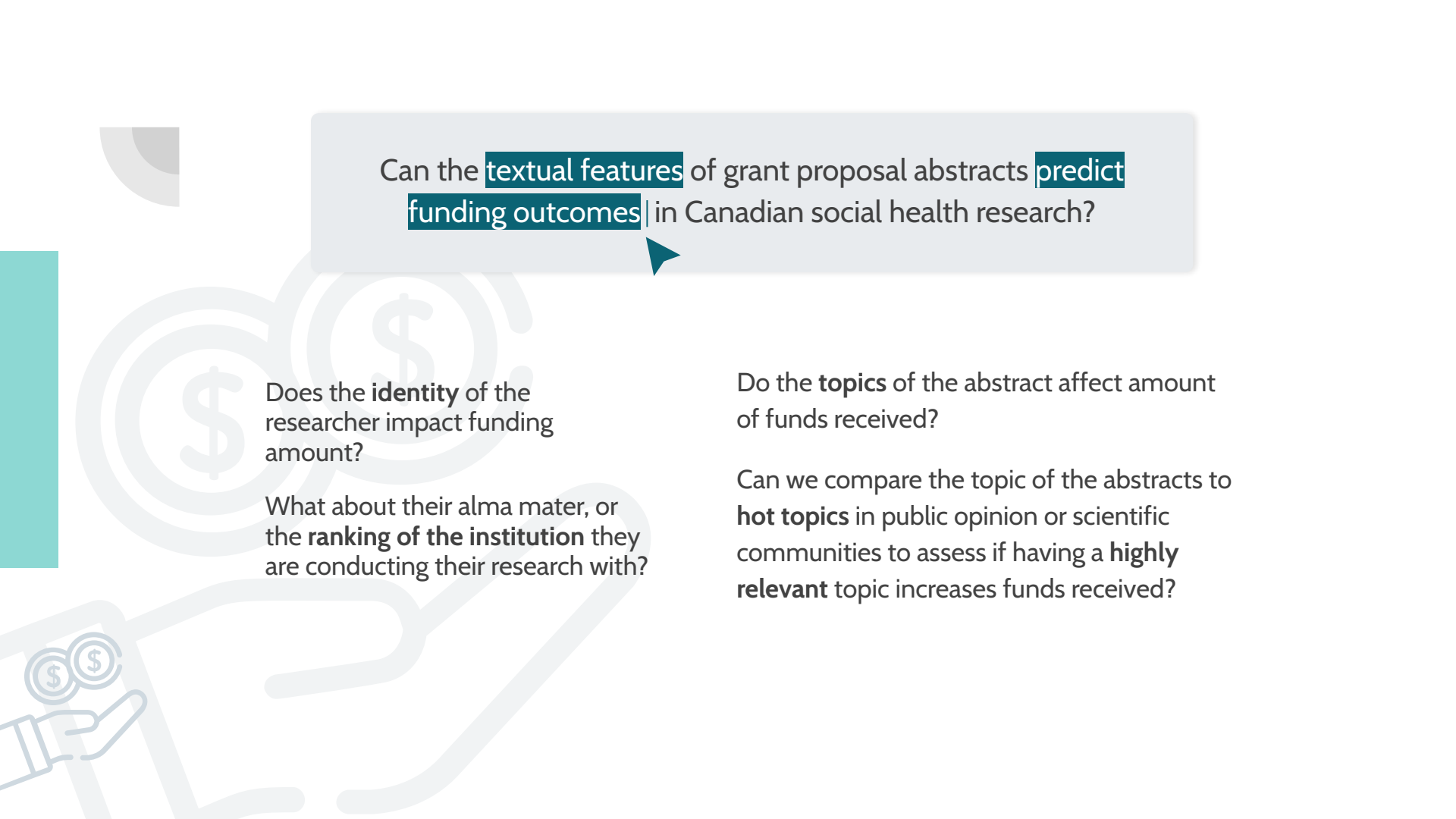
- significantly reduce time spent reviewing grants
- reduce human bias in the evaluation process

Thus, **our research can make the grant review process more objective, efficient, and transparent.**



Research Question





Can the **textual features** of grant proposal abstracts **predict funding outcomes** in Canadian social health research?

Does the **identity** of the researcher impact funding amount?

What about their alma mater, or the **ranking of the institution** they are conducting their research with?

Do the **topics** of the abstract affect amount of funds received?

Can we compare the topic of the abstracts to **hot topics** in public opinion or scientific communities to assess if having a **highly relevant** topic increases funds received?

Data Source & Methods



Data Source

Our dataset is exported on November 17th from the Government of Canada's Canadian Institutes of Health Research funding Decision Database.

Focus Theme: Social/Cultural/Environmental/Population health

Geographic Scope: USA, Canada

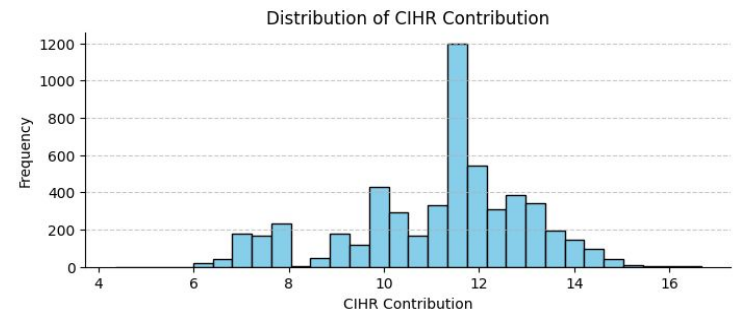
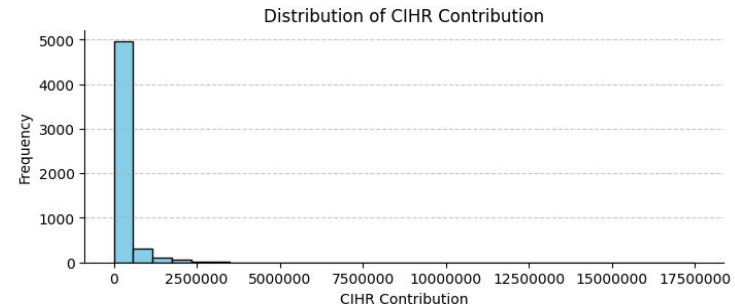
Total projects: 7011 projects

Key Variables:

- Institution Paid – Institution responsible for the project (categorical)
- Abstracts – Abstracts of the proposed projects (text)
- **CIHR Contribution** – The funding amount provided by the Canadian Institutes of Health Research (numerical)

Descriptive Statistics:

- No missing data
- Around 1% of the total are duplicates



Source:

<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%20country%3A%22United%20States%20of%20America%22&sort=namesort%20asc&start=0&rows=20>



Methodology

Topic Modeling (LDA)

Extract latent themes for abstracts

Train-Test Split

Ensure fair evaluation on unseen data

Model Evaluations

Select the best performing model for our project



Data Processing

Clean text to ensure meaningful topic extraction

Feature Engineering

Incorporate novelty score, project term years and text features to enrich the dataset

Model Training

Test multiple algorithms to assess predictive power

Data Processing and Topic Modeling

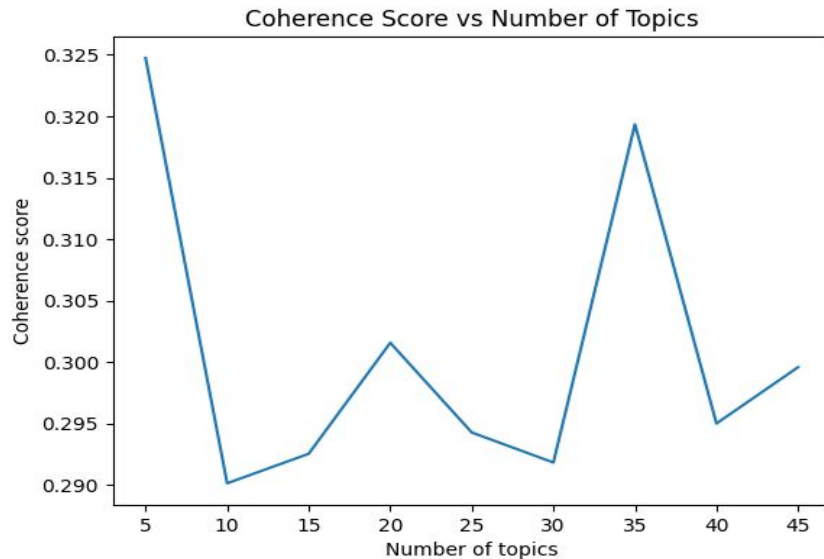
1. Text Preprocessing:

- Scraped data using **selenium**
- Tokenized and cleaned abstracts
- Extracted bigrams using **gensim**



2. Topic Modeling:

- Used **Latent Dirichlet Allocation (LDA)** to identify latent themes across abstracts
- Optimal number of topics (**5**) determined using coherence scores
- Generated topic distributions for each abstract





Feature Engineering and Train-Test Split

3. Novelty Scores:

- Measured how novel each abstract is compared to other abstracts using cosine similarity.
- Scores range from 0 - 1, reflecting textual uniqueness.

4. Train-Test Split and Scaling:

- Split data into 80% training and 20% testing.
- Scaled numerical features using StandardScaler.





Model Training and Evaluation

6.1. Algorithms Used:

- Linear Models: Linear Regression, Lasso, Ridge, Elastic Net
- Tree-Based Models: Decision Tree, Random Forest, Gradient Boosting
- KNN

6.2. Hyperparameter Tuning:

- Used GridSearchCV with 5-fold cross-validation to tune model hyperparameters
- Extracted the best performing models

7. Model Evaluation Metrics:

- Mean Squared Error (MSE): measures average squared prediction error
- R^2 Score: measure variance explained by the model



Literature Review

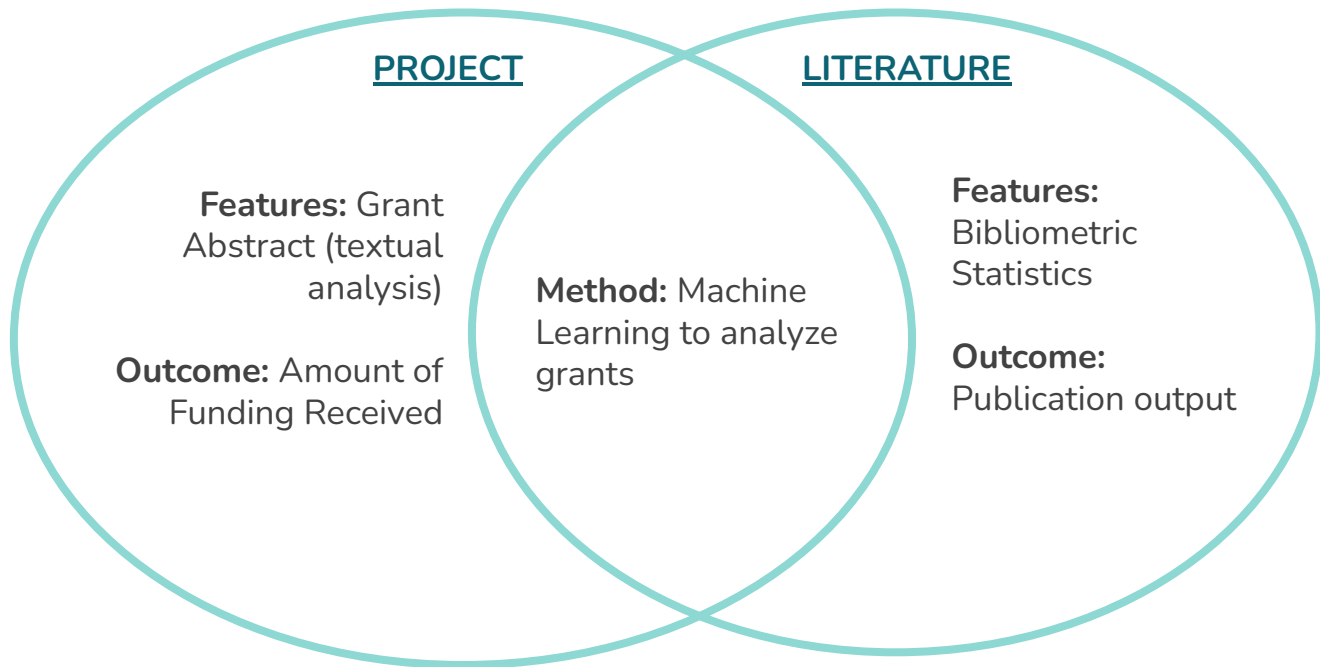




On Predicting Research Grants' Productivity Via Machine Learning - Tohalino and Amancio

ABOUT:

Tohalino and Amancio employ machine learning to **predict grant productivity** for grants in medicine, dentistry, and veterinary medicine. The article finds that the **topic and year of publication** to be the most significant predictors.



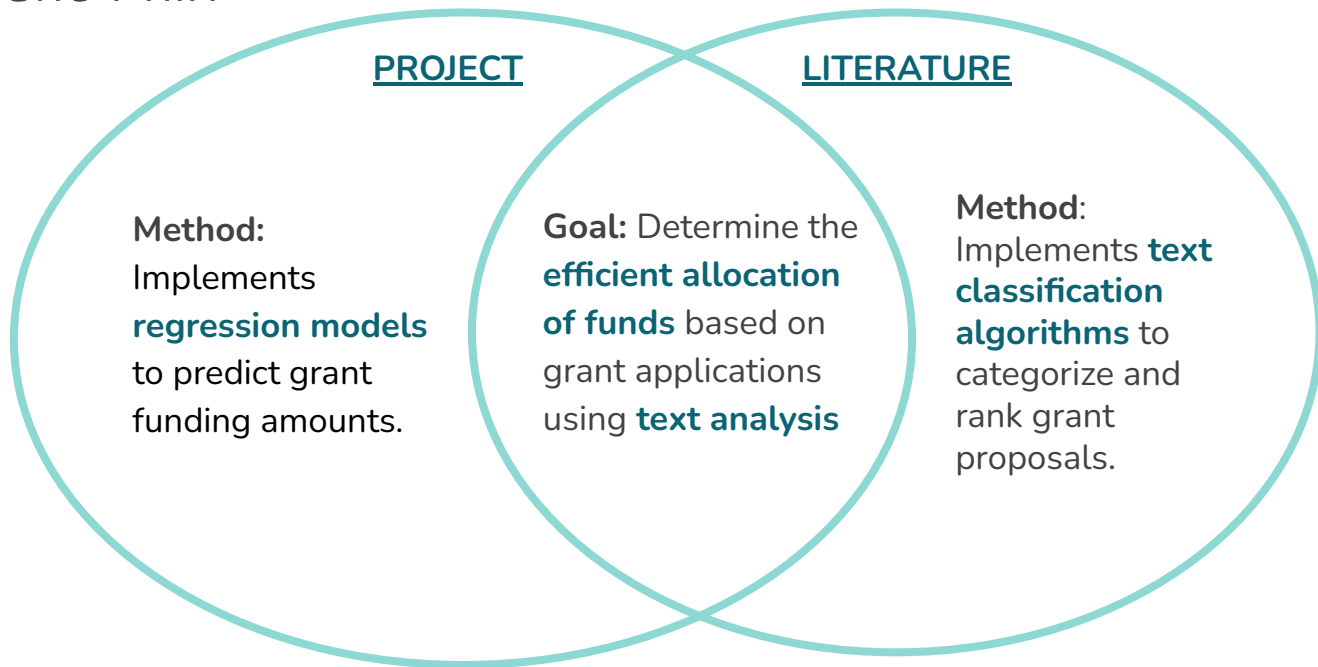


From Algorithms to Grants: Leveraging Machine Learning for Research and Innovation Fund Allocation

- Lupyani and Phiri

ABOUT:

The paper explores the application of **machine learning** to improve the grant allocation process. By analyzing historical grant data and employing **text classification models**, the study aims to automate the evaluation of proposals, and increase the transparency of funding decisions.



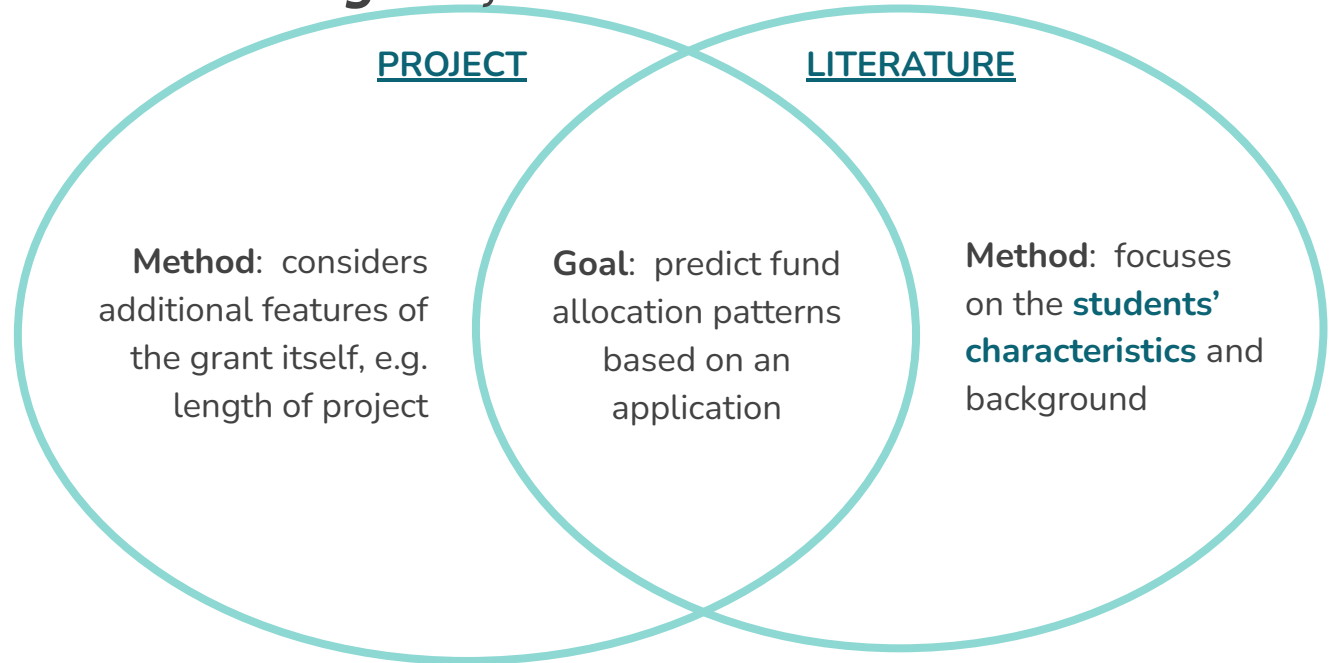


A Data-Driven Approach in Predicting Scholarship Grants of a Local Government Unit in the Philippines Using Machine Learning - Fajardo et al.

ABOUT:

Fajardo et al. develops a machine learning model to match **scholarship applicants** in the Philippines with the best scholarship.

The article finds the **logistic regression** to be the best-performing model.



Results





Topic Modelling

The optimal number of topics was 5, with a coherence score of 0.325.



Topic 1 - Women's Health in Canada

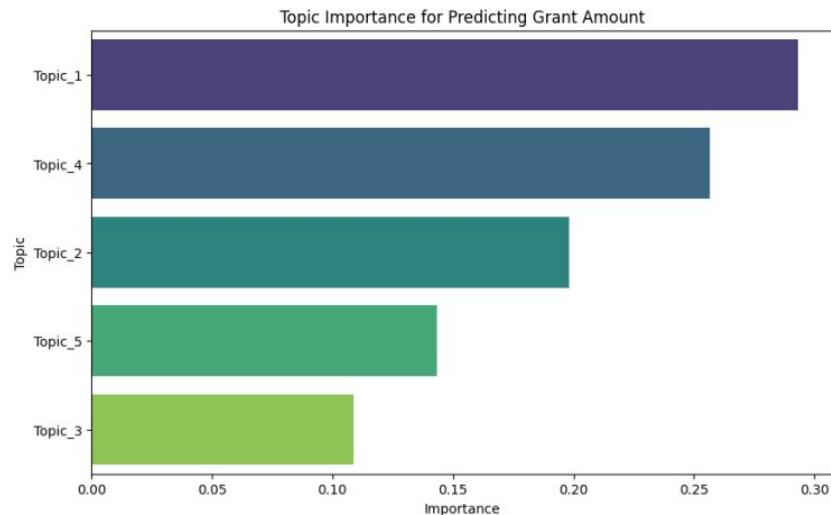
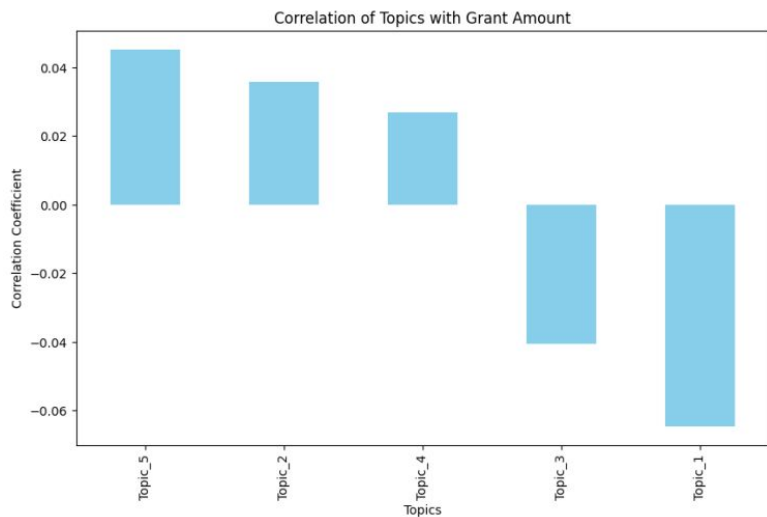
Topic 2 - Youth-Centered Policy

Topic 3 - Mental Health and Social Outcomes

Topic 4 - Indigenous Community Health

Topic 5 - HIV Risk and Canadian Population Health

Correlation: Topics and Grant Amounts



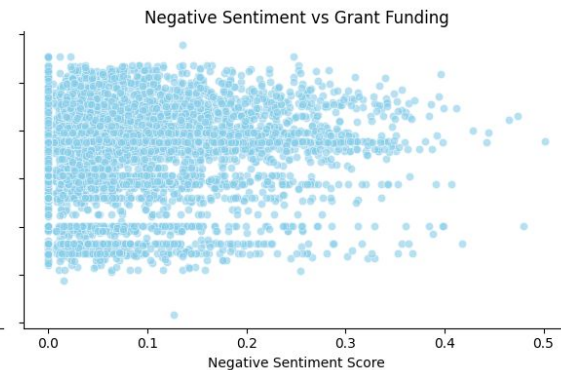
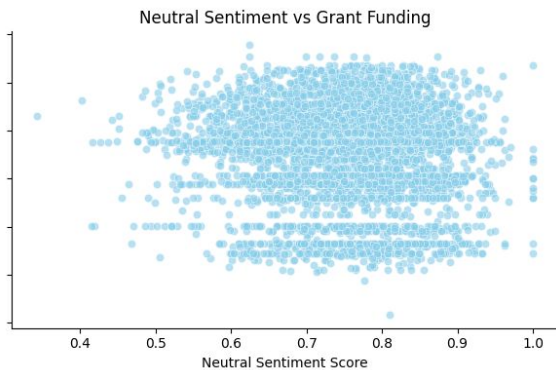
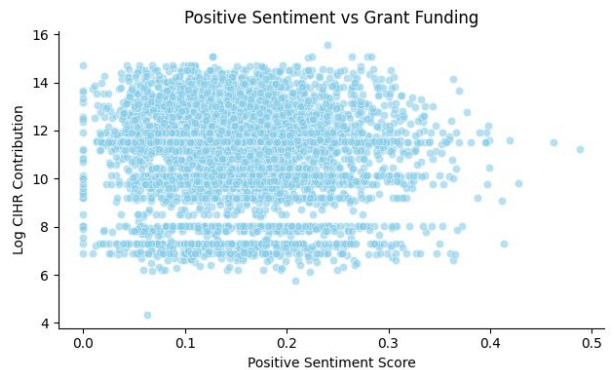
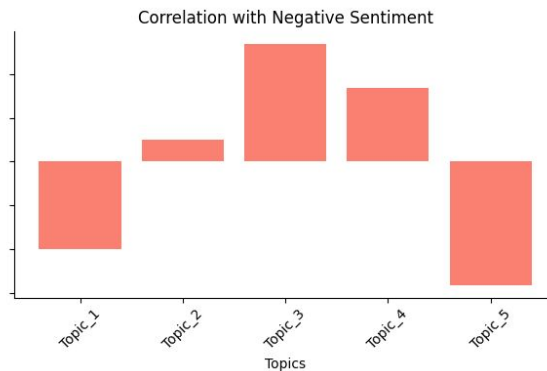
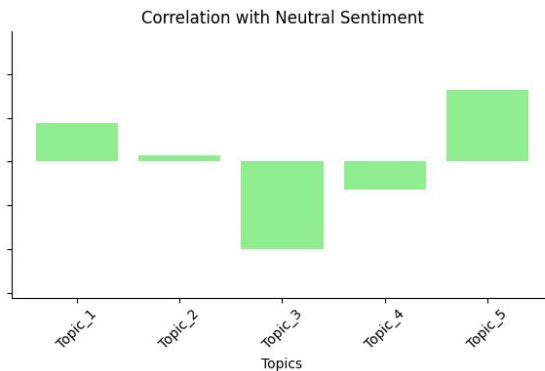
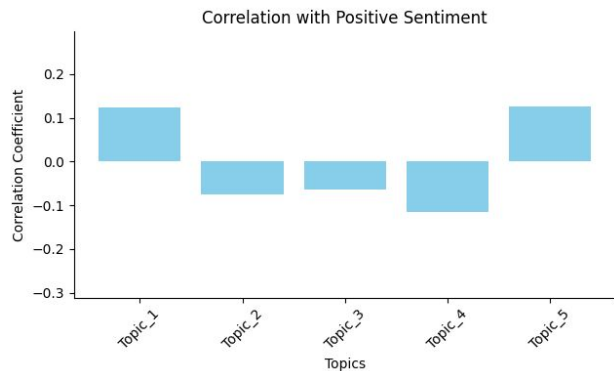
Topic 1 (Women's Health in Canada) is the **most important feature** for predicting grant amounts - with a **negative correlation**. Topic 1 **most strongly predicts lower grant amounts**.

Topic 4 (Indigenous Community Health) is the next most important feature, and is moderately positively correlated with grant amount, being **predictive of and associated with higher grant amounts**.

Topic 3 (Mental Health and Social Outcomes) is the **least predictive** of grant amount and is **negatively correlated** with grant amounts.

Sentiment Analysis

- VADER lexicon
- Less variation in positive/neutral, more in negative



Model Evaluation: Grant Amount Regression

Hyper Parameters:

- Trees: 100
- Learning Rate: 0.1
- Max Depth: 5

Features:

- Topic distribution (from LDA)
- Term years (length of project)
- Year (year which grant was submitted)
- Institution (dummy)
- Abstract length
- Novelty score (uniqueness of abstracts)

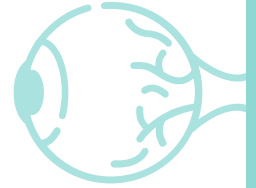
MODEL	R ²	MSE
Linear Regression	-1.0703 e+22	9.6195 e+22
Ridge Regression	0.4460	1.8736
Lasso Regression	0.4517	1.8543
Elastic Net	0.4517	1.8543
Decision Tree	0.5241	1.6095
Random Forest	0.5590	1.4916
Gradient Boosting	0.5714	1.4496
KNN	0.3312	2.2618

Conclusion





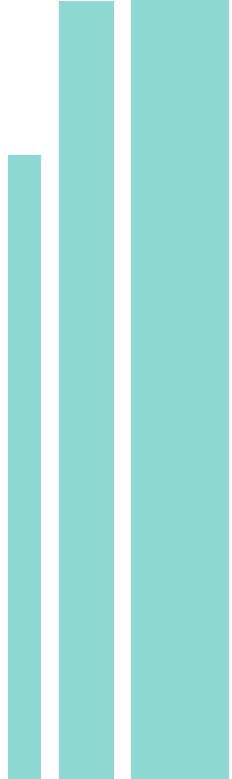
Limitations



Area of Focus: We focused on Social / Cultural / Environmental / Population Health theme from the CIHR database, leaving **other themes** in Health Research such as Biomedical or Clinical research **unexplored**

Monetary Constraints: More optimal resources were locked behind a paywall, e.g. the LIWC lexicon, suitable for **academic content**.

Limited Variables: According to the 57% R^2 score, there is still 43% of variation in grant allocation unexplained. **Characteristics and background of the grant review committee** were not taken into account, and neither were **qualifications or demographics of the applicants**.

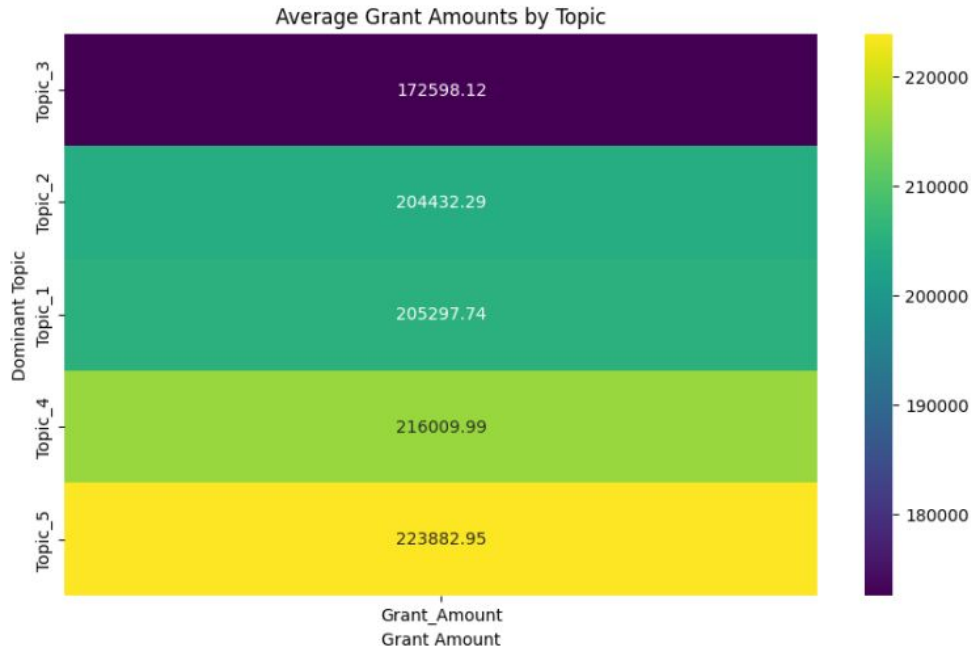


Identifying Bias

Topic 1 (Women's Health in Canada) was the most important predictor of grant allocation and had the **largest correlation**, which was **negative**

Topic 3 (Mental Health and Social Outcomes) received the **least funding**. It was not very predictive but did have a notable **negative correlation**.

This suggest that grant review committees may have viewed these topics as **not as worthy of funding**, revealing **potential bias** against marginalized groups in healthcare such as **women** and those struggling with **mental health**.



A large, faint, light blue background graphic on the right side of the slide. It depicts a human brain with several interlocking gears of different sizes, symbolizing research, thought, and innovation. The brain is oriented vertically, with the top of the head towards the top of the slide.

Further Research

Topic Relevance: Relation of grant abstract to current **hot-topics in public opinion** or scientific fields would be an interesting and relevant predictor of grant allocation.

Applicant Qualifications: Does the **applicant's research history** (renown, achievements, past publications) impact grant allocation?

Productivity of Grant: While we focused on the monetary grant received by applicants, it would be prudent to investigate the **results of the funded research** - this can help determine if grants are being efficiently allocated.



References

- 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>
- 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>
- 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>
- Icons by Flaticon



Thank You! Questions?