Ministry of Higher Education,
Research & Innovation

Report for Funded Research Grant

Project Title:

Using Text and Data Mining (TDM) techniques to improve the mechanisms for evaluating research proposals - a future study to meet the Goals of Oman 2040 vision

Project ID:

MOHERI - BFP/GRG/ICT/21/052

Submitted by:

PI : Sheikha Obaid Al Mujaini

Sultan Qaboos University

almujaini@squ.edu.om

May, 2021

Table of Contents

List of Figures

List of Tables

1. Executive Summary

In the era of the information revolution that we are experiencing during the 21st century, the need for mechanisms for sorting and mining texts and information has become of utmost importance to ensure benefit from the information available in various information sources. The techniques used for data and text mining are among the most widely used modern techniques in various sectors, starting from the education sector to the industrial and economic sectors, which have been widely employed within various fields and have proven their effectiveness. In view of the research sector, whose research fields vary based on the national need, the strategic directions and plans of institutions, in addition to the research interests of researchers, there is an urgent need to benefit from these mechanisms and techniques to improve the process of selecting competing projects to obtain research funding to be in line with the rapid scientific development and the future need of the market and the external community. By reviewing the various research papers that have used these technologies for several different goals, according to the aspirations of those studies, this study will analyze the current standards used at Sultan Qaboos University to evaluate and sort competing research projects. It will also design an automated information system to search available texts and information, which will automatically sort competing projects according to the basic standards required by the funding body first, and the research standards on which Sultan Qaboos University's research strategy, and finally the pillars of Oman Vision 2040. Accordingly, the level of efficiency of the projects funded for Sultan Qaboos University researchers will raise, so that they will be compatible with the future vision of the Sultanate of Oman. On the other hand, the funding opportunities for projects with similar topics will reduce.

Keywords: competing projects, text and information mining mechanisms, Oman Vision 2040.

● INTRODUCTION

In the era of the modern information revolution and the great information momentum distributed in various information sources, the process of sorting and mining data has become one of the stressful processes. Accordingly, the need has emerged to use modern technology to sort, arrange and employ information and benefit from it to answer various research questions and use it to develop future plans and visions in various fields. For various purposes, mechanisms for analyzing and mining information and texts will enable researchers in the humanities and scientific research fields to build their research on the latest research results reached by researchers in the same field, which will push the cause of research development forward (San Tay, P., & Sik, C.P. 2016).

Text and data mining mechanisms have been widely used in various fields such as medicine, law, and commerce, in addition to the security field (Truyens, & Van Eecke, 2014). These mechanisms have been used to analyze customer behavior and purchasing tendencies and determine the market need based on that. They have also been used To prevent crime and combat terrorism. As for its applications in the field of biomedicine, it has been used to develop treatments for various diseases and future forecasts for epidemics (Johnson, R., Fernholz, O., Fosci, M. (2016)), and such technologies have helped research institutions know the research directions of researchers. And redirect it according to future plans and the need for that, thanks to its ability to shed light on the latest findings of researchers in this field, the shortcomings, and what is the future direction of research and studies (Anjewierden, A., Kolloffel, B., & Hulshof, C. 2007).

There are two types of mining processes: Text Mining (TM) and Data Mining (DM). The Data Mining (DM) process relies on organized and homogeneous data that you usually find in various databases. However, the Text Mining (TM) process It is based on data available, for example, in documents and social networking sites, which are heterogeneous and must be arranged before starting the analysis process. Note that most of the data available globally exists in an uncoordinated and homogeneous form, which represents 80% of the data in the world (Dang, S., & Ahmad , P. H. (2014).) Those in charge of mining techniques may be required to combine text mining mechanisms and data mining mechanisms so that they can obtain richer information. There are many techniques that you may be exposed to in the process of data analysis, such as statistical processes, artificial intelligence techniques (AI) and machine learning (ML) processes, as they are complementary sciences for obtaining a mixture of information for different research purposes. (OpenText Blog, 2019).

The mechanics of implementing text and data mining (TDM) techniques vary depending on what they are used for, but they usually share general basic steps. This difference is due to several factors, including the software used to do the mining, in addition to the experience of the analyst performing the task. The complexity of the data used and the purpose behind doing the mining are also factors that determine the stages of implementing these mechanisms. The clearer the objectives behind carrying out the excavation process are, the more the desired objectives will be achieved (San Tay, P., & Sik, C. P. 2016). Most text and data mining mechanisms usually consist of four different stages: knowing the quality of the entered data, whether it is structured and homogeneous data or vice versa, and accessing it. The second stage consists of copying large quantities of materials and preparing them so that they are machine readable and can be used to extract the required information. Third. In the fourth and final stage, the data is reassembled to identify new patterns and discover new knowledge through the final output of the mining process (Geiger, C., Frosio, G., & Bulayenko, O. (2018)).

Research projects and consultations are considered important pillars of research institutions that bring them tangible and intangible benefits, as these educational and research institutions constitute a house of expertise and an environment for diverse and modern experiments and studies that may provide solutions to existing problems in different environmental, industrial, economic and agricultural sectors. Hence the need to improve mechanisms for evaluating and selecting research projects in order to achieve the aspirations of educational institutions and researchers and also provide practical, applied solutions that can be used by the external community. Hence the need to improve mechanisms for evaluating and selecting research projects in order to achieve the aspirations of educational institutions and researchers and also provide practical, applied solutions that can be used by the external community. And based on the royal directives of His Majesty Sultan Qaboos bin Said, may God have mercy on him - (at that time) towards drawing a future vision for Oman during the next twenty years, which draws a map of the nation that accommodates the economic and social reality and looks forward to the future. This vision was ratified by His Majesty Sultan Haitham bin Tariq in December 2020, as the work of the sectoral committees and awareness of their importance at the internal and external levels began in March 2017 and was completed and announced in the first quarter of 2019. The vision focuses on four main themes: human and society, economy and development, governance and institutional performance, and the sustainable environment, twelve pillars fall under each theme, which contribute to its enrichment and development. Several basic starting points were relied upon in the preparation phase of the future vision, namely: the national priorities of the Sultanate, the report on the main guidelines for formulating the future vision of Oman 2040, as well as the national program for promoting economic diversification “Tanfeedh”, the outputs of the Vision 2040 committees and work teams, strategic studies and reports, and the lessons and achievements of the vision. Oman 2020, the 2030 Sustainable Development Goals issued by the United Nations, sectoral strategies, the National Urban Development Strategy, the Ninth Five-Year Development Plan, international reports and indicators related to the pillars of the vision, outputs of the Vision 2040 Office (Oman Vision 2040, n.d.).

Accordingly, and in line with the national trend towards achieving the goals of the vision during the next twenty years, the educational and research institutions in the Sultanate, which is considered one of the main building blocks under whose roof experts and specialists in various scientific fields remain, are an important pillar for achieving the set goals, as the total national spending on research and development reached A percentage of 0.22 of the gross domestic product for the year 2018 AD, according to indicators recorded in the National Bank (World Bank, P.T.). Therefore, working on evaluating the mechanisms for selecting funded research projects is a sensitive and important process for sorting competing projects based on national priorities and the pillars of the future vision of Oman in order to ensure the proper distribution of research budgets. In view of the directives of Future Vision 2040, self-sufficiency and reliance on domestic product, such as the agricultural and fisheries sector, are an important contributor to the Omani economy, and not being completely dependent on oil and gas resources as a main driver of the Omani economy, despite achieving 68%-85% of the total average government revenues, in addition to Reliance on the tourism sector as one of the important national economic factors. Accordingly, research is considered an important source of expertise in the areas that Oman Vision 2040 focuses on (Nordea, 2020).

This study will analyze the mechanisms used at Sultan Qaboos University to select competing projects to obtain research funding, and the extent of their effectiveness in the proper selection of projects in line with the university's research strategy and the standards set by the funding body, in addition to the pillars of Oman Vision 2040. It will also apply text and data mining mechanisms to carry out the selection and sorting process, thus ensuring that projects of national value and direction in line with the future vision are awarded, and reducing the chances of wasting research funding on projects with repetitive content or those that do not provide society with the required need in all sectors.

● OBJECTIVES

This study aims to obtain several results, which are as follows:

1. Knowing the university's researchers' research orientations and the extent to which this research is compatible and consistent with the Oman vision 2040 and thus the university's first fifth vision (2021 to 2025).
2. Establish a screening mechanism for competing research to obtain funding according to the themes of Oman Vision 2040.
3. Improving the criteria for evaluating competing research approved by the Deanship of Research and amending them based on the extracted results in order to ensure their consistency with Oman's future vision, and to ensure the proper distribution of research budgets in addition to increasing the effectiveness of benefiting from the results of research projects in the local and global community.
4. Increase the opportunity to fund research with a new idea and good content.
5. Evaluating the extent to which regular paper-based evaluation is similar to evaluation using text mining techniques, in order to study the effectiveness of relying on modern techniques to reduce the effort and time used to evaluate scientific research.

● MATERIALS AND METHODS

This section details the systematic approach and methodologies applied in developing a classification model for evaluating and classifying research proposals. Our process involves several distinct stages, including data preparation, feature extraction, model training, evaluation, and application to real-time data. The methodologies are underpinned by machine learning principles and natural language processing techniques.

1. Data Preparation:

   o Source Materials: The dataset comprises research proposals stored in two formats: PDF and JSON. These proposals are categorized into two classes: 'Approved' and 'Rejected'. We chose the test data for three types research funding programs, Internal research grants (IG), His Majesty research grants (SR) and Ministry of Higher education, research and Innovation' research (MoHERI) grants (RC) over the period from 2018 to 2024.

Type	Page Number	Words number	Approved	Rejected	Filed Number
IG	44.5	20,588,294	394	63	457
RC	10.4	4,377,652	60	150	210
SR	64.7	8,000,347	19	87	106

- Extraction and Preprocessing:
  - ✓ Text was extracted from PDF documents using PyPDF2, a Python library.
  - ✓ JSON files were parsed to extract relevant textual content.
  - ✓ Text preprocessing involved cleaning, normalization, and tokenization to prepare the data for feature extraction.

2. Feature Extraction:

- TfidfVectorizer: statistical formula to convert text documents into vectors based on the relevancy of the word. It is based on the bag of the words model to create a matrix containing the information about less relevant and most relevant words in the document.
- Vectorization: We employed TfidfVectorizer from the Scikit-learn library to convert textual data into a structured, feature-rich format suitable for machine learning models.
  - ✓ The vectorization process creates a numerical representation of the text by evaluating the importance (frequency) of specific words or phrases within the dataset compared to their occurrence across all documents.

3. Model Development and Training:

Rationale for Choosing the Classification Model:

After careful consideration, we decided to use the classification model for the following reasons:

- **Nature of the Problem:**
  - ✓ The problem at hand is inherently a classification problem. We need a clear decision on whether a proposal should be approved or rejected, which aligns directly with the capabilities of a classification model.
- **Effectiveness:**
  - ✓ Classification models, especially RandomForestClassifier, are highly effective in handling complex decision-making tasks. They can capture intricate patterns in the data and provide robust predictions.

- **Handling Imbalance:**
  - ✓ Our dataset exhibited a significant imbalance between the number of 'Approved' and 'Rejected' proposals. The ADASYN technique, used in conjunction with the classification model, helps address this imbalance by generating synthetic samples for the minority class, ensuring fairer and more accurate predictions.
- **Comprehensive Evaluation:**
  - ✓ The classification model allows for a comprehensive evaluation using metrics like precision, recall, and F1-score, which provide a detailed understanding of the model's performance in correctly identifying 'Approved' and 'Rejected' proposals.

- **Decision-making Insights:**
  - ✓ While similarity calculations can indicate alignment with criteria, they do not provide a binary decision. The classification model directly answers the critical question: "Should this proposal be approved or rejected?" This binary decision-making is more practical and actionable for committee evaluations.

- Classification Algorithm: A RandomForestClassifier was chosen for its effectiveness in handling complex classification tasks and its inherent capability to manage overfitting.
- Balancing Technique: Given the imbalance observed between 'Approved' and 'Rejected' classes due to the high number of rejected research proposals compared to the approved ones, we applied the ADASYN (Adaptive Synthetic Sampling Approach) technique to generate synthetic samples for the minority class, thus balancing the dataset before model training.
- Pipeline Construction: A pipeline combining the TF-IDF vectorization, ADASYN oversampling, and the RandomForest classifier was constructed to streamline the training process.

4. Model Evaluation:

- Cross-Validation:
  - ✓ is a technique used in machine learning to evaluate the performance of a model on unseen data.
- Performance Metrics
  - ✓ Precision:

    Precision measures the proportion of true positive (TP) predictions among all positive predictions made by the model. It is calculated as the ratio of TP to the sum of TP and false positives (FP).
  - ✓ Recall

    Recall, also known as sensitivity or true positive rate, measures the proportion of true positive predictions among all actual positive instances. It is calculated as the ratio of TP to the sum of TP and false negatives (FN).

- ✓ F1-score

  F1 Score is a metric that balances precision and recall. It is calculated as the harmonic mean of precision and recall. F1 Score is useful when seeking a balance between high precision and high recall, as it penalizes extreme negative values of either component.

- o Our model's performance was assessed using precision, recall, and F1-score metrics, particularly focusing on its ability to correctly identify proposals from the minority 'Approved' class.
- o Cross-Validation: To ensure the model's robustness and generalizability, 5-fold cross-validation was employed, calculating weighted averages for the F1-score across different subsets of the dataset.

5. Real-time Classification:

At this stage we have used new, unclassified data (Not approved nor rejected). To test the model on new data, we have done this:

- o Script Development: A Python script was developed to classify new research proposals in real time. The script automatically detects the file format, extracts text, and applies the trained model to output a classification score and decision.
- o Usage: Will enter the filename of a new proposal into the script, which then returns the model's classification ('Approved' or 'Rejected') along with a confidence score.

6. Visual Aids and Notation:

- o precision-recall curve
  - ✓ shows the tradeoff between precision and recall for different thresholds.
- o Figures illustrating the model's precision-recall curve and other relevant statistics were generated to provide visual insights into the model's performance.
- o Tables summarizing the classification report and cross-validation results were included to present quantitative evaluation metrics in a clear and concise format.
- o Equations, such as the formula used for calculating the F1-score, were clearly numbered and formatted for easy reference. For example, the F1-score is calculated as follows:

$$F1 = 2 \times precision + recall / precision \times recall (1)$$

Similarity Analysis using Multilingual BERT:

At this stage We have applied the twelve strategic national priorities of Oman 2024 vision over the approved proposals only of the three types of grants we are testing (IG, SR & RC). The below techniques were used:

- Embedding Extraction:
    - Multilingual BERT was used to generate embeddings for both the proposals and the predefined strategic national priorities. These embeddings capture the semantic meaning of the texts, facilitating accurate similarity comparisons.
- Cosine Similarity Calculation:
    - Cosine similarity was employed to measure the alignment between proposal embeddings and strategic national priorities embeddings. This similarity measure helps determine how well a proposal aligns with key areas of focus such as Oman Vision 2040 and SQU focus areas. (See Image 3)

$$\cos(A, B) = \frac{AB}{|A||B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2}\sqrt{\sum_{i=1}^{n} (B)^2}}$$

Real-time Classification and Similarity Evaluation:

- Script Development:

    - A Python script was developed to classify new research proposals in real time. The script automatically detects the file format, extracts text, and applies the trained model to output a classification score and decision.
- Similarity Scoring:
    - The script also calculates similarity scores between the proposal text and strategic national priorities, providing additional context on how well the proposal aligns with predefined goals.

• RESULTS AND DISCUSSIONS

This section presents a comprehensive analysis of the machine learning model designed to classify research proposals into 'Approved' and 'Rejected'. The performance assessment is based on the model's ability to accurately predict these categories across different datasets.

Dataset Overview:

- Data Composition: The dataset included proposals categorized under internal Grant (IG) and Strategic Grants (SR), and Innovation MoHERI grants (RG).
- Data Split: Training (80%) and Testing (20%).
    - we split the data to allows us to train a model on the training set and then test its accuracy on the testing set

Performance Metrics:

- Accuracy: The overall accuracy of the model was 87%, reflecting its ability to correctly classify both 'Approved' and 'Rejected' proposals on unseen data.

Precision and Recall:

- Precision for Approved Proposals: 88%
    - Explanation: This figure means that when our system predicts that a proposal should be approved, it is correct 88% of the time. In simpler terms, if our system gives a green light to 100 proposals, 88 of them truly deserve approval based on the criteria.
- Recall for Approved Proposals: 92%
    - Explanation: This metric shows how well the system is at catching all the proposals that truly deserve approval. A 92% recall rate means that out of all the proposals that should be

approved, our system successfully identifies 92 out of every 100. This ensures that very few good proposals are overlooked, minimizing the chance of missing out on worthy projects.

- Precision for Rejected Proposals: 86%
    - Explanation: This indicates the accuracy of the model in identifying proposals that should be rejected. An 86% precision rate means that when our system rejects a proposal, it is correct 86% of the time. So, for every 100 proposals it rejects, 86 of them genuinely do not meet the required standards.

- Recall for Rejected Proposals: 80%
    - Explanation: Recall for rejected proposals measures the system’s ability to identify all the proposals that should not be approved. An 80% recall rate indicates that the system identifies 80 out of every 100 proposals that should be rejected. This suggests that 20 out of every 100 proposals that deserve rejection might slip through, representing areas where the system could improve to catch more of these unsuitable proposals.

F1-Score:

- Approved: 90%, indicating a strong balance between precision and recall for approved proposals.

Detailed Classification Performance

Classification	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Approved	88	92	90
Rejected	86	80	83

Visualizations:

- Figure 1: Precision-Recall Curve — Illustrates the trade-offs between precision and recall, highlighting the model's ability to balance these metrics at different thresholds.

7. Evaluation on New Data

The RandomForestClassifier was applied to a new set of unseen data to assess its performance in a real-world scenario. The classification report for this evaluation reveals significant disparities in performance between the 'Approved' and 'Rejected' categories:

Precision for 'Rejected' (60%)

- **Explanation**: Precision tells us how many of the model's 'Rejected' decisions were actually correct. With a precision of 60%, it means that out of every 10 proposals the model rejects, about 6 are rightly rejected. This is good but not great, indicating that the model sometimes labels proposals as 'Rejected' when they should not be.

Recall for 'Rejected' (12%)

- **Explanation**: Recall measures whether the model is good at catching all the cases it needs to catch. A recall of 12% for 'Rejected' proposals is very low. It means that out of every 100 proposals that should be rejected, the model only identifies 12 correctly. This suggests the model is overlooking many proposals that it should reject.

F1-Score for 'Rejected' (20%)

- **Explanation**: The F1-score helps balance the views given by precision and recall. An F1-score of 20% is quite low, which tells us that the balance between catching rejections accurately (precision) and catching enough of them (recall) is not very good. The model is neither accurate enough nor comprehensive enough in identifying proposals that should be rejected.

Precision for 'Approved' (84%)

- **Explanation**: This tells us that when the model approves a proposal, it is correct about 84% of the time. This high precision indicates that the model is quite reliable in approving proposals that indeed meet the criteria.

Recall for 'Approved' (98%)

- **Explanation**: This shows that the model is excellent at identifying proposals that should be approved, catching 98 out of every 100 such cases. This high recall rate means very few good proposals are mistakenly rejected.

F1-Score for 'Approved' (90%)

- **Explanation**: A high F1-score of 90% for approved proposals signifies a strong balance between the model's precision and recall in this category. It indicates that the model is both accurate and effective in approving the right proposals.

Overall Model Accuracy (82%)

- **Explanation**: Overall accuracy tells us how many of the model's decisions were correct out of all decisions made. An accuracy of 82% suggests that the model performs well overall. However, this metric might not fully highlight the model's struggles with the 'Rejected' category.

Classification	Precision (%)	Recall (%)	F1-Score (%)
Rejected	60	12	20
Approved	84	98	90

8. Real-time Classification and Similarity Analysis

In this section, we introduce the methodology employed for real-time classification and similarity analysis of research proposals. This process leverages the power of Multilingual BERT to ensure that the proposals are not only evaluated for approval or rejection but also assessed for their alignment with the strategic national priorities outlined in Oman Vision 2040 and SQU focus areas.

Embedding Extraction with Multilingual BERT:

2. **Model Selection:**

- We utilized the Multilingual BERT model from the Hugging Face Transformers library. This model is pre-trained on a large corpus of multilingual data, making it suitable for our text analysis tasks which include Arabic content.

3. **Text Processing:**
    - Text data from proposals are tokenized and processed to fit the input requirements of the BERT model. Tokenization includes splitting the text into tokens and adding special tokens required by the BERT model..
4. **Embedding Generation:**
    - The processed text is passed through the BERT model to obtain embeddings. Specifically, we use the pooler_output of the BERT model, which provides a fixed-size vector representation of the entire input sequence.
    - These embeddings capture the semantic meaning of the proposals, facilitating accurate similarity comparisons.

Similarity Calculation:

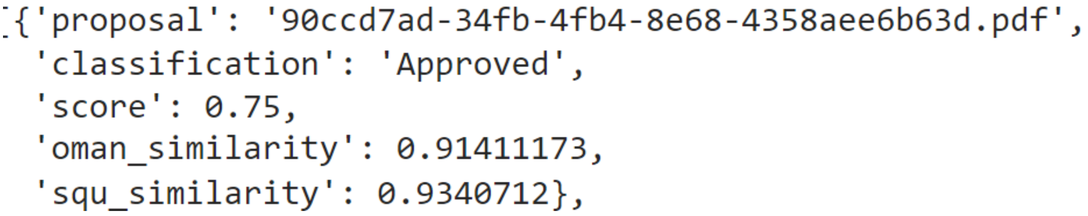
1. **Strategic national priorities Embeddings:**
    - Predefined texts representing strategic national priorities from Oman Vision 2040 and SQU focus areas are similarly processed and converted into embeddings using the same BERT model. This step ensures consistency in the representation space.
2. **Cosine Similarity:**
    - Cosine similarity is employed to measure the alignment between the proposal embeddings and the strategic national priorities embeddings. The similarity scores range from -1 to 1, where a higher score indicates greater alignment.
    - The similarity calculation is performed for each strategic priority to identify which proposals align most closely with the key areas of focus.

Implementation:

1. **Real-time Classification Script:**
    - A Python script was developed to automate the entire process. The script performs the following functions:
        - **Text Extraction:** Extracts text from PDF and JSON files using PyPDF2 and JSON parsing libraries.

    - ■ **Classification:** Uses a pre-trained RandomForestClassifier to classify proposals as 'Approved' or 'Rejected'.
    - ■ **Embedding Calculation:** Computes embeddings for the extracted text using the Multilingual BERT model.
    - ■ **Similarity Analysis:** Calculates cosine similarity scores between the proposal embeddings and strategic national priorities embeddings.
    - ■ **Output:** Prints the classification results along with similarity scores in real-time for each processed proposal.

- **Sample of the output**:



```
{'proposal': '90ccd7ad-34fb-4fb4-8e68-4358aee6b63d.pdf',
 'classification': 'Approved',
 'score': 0.75,
 'oman_similarity': 0.91411173,
 'squ_similarity': 0.9340712},
```

75% indicate the model is sure 75% that the proposal is approved.
91.4 indicate that this proposal is similar 91.4% to Oman vision 2040
93.4 indicate that this proposal is similar 93.4% to Sultan Qaboos university criteria.

2. **Handling Different File Formats:**
    - ○ The script is designed to handle both PDF and JSON files seamlessly, ensuring that text extraction is robust and accurate regardless of the file format.

Evaluation:

1. **Classification and Similarity Output:**
    - ○ For each proposal, the script outputs the classification ('Approved' or 'Rejected') along with the corresponding probability score.
    - ○ Additionally, the script provides similarity scores indicating how well the proposal aligns with each strategic priority from Oman Vision 2040 and SQU focus areas.

i. Discussion

In this project, we implemented the ADASYN technique to balance our dataset. Balancing the dataset is crucial in classification problems because imbalances between classes can lead to biased results, especially when the data skews heavily toward one class. also developing a robust classification model, and integrating real-time similarity analysis using Multilingual BERT.

- **Importance of Balancing Techniques:**

Using ADASYN helps create synthetic samples of the underrepresented class ('Rejected' proposals in our case) to equalize the influence of both classes during the model's training process. This approach aims to provide a more generalizable and fair decision-making model.

- **Challenges in Identifying 'Rejected' Proposals:**

Our analysis revealed significant challenges in correctly identifying 'Rejected' proposals. The model exhibited a very conservative behavior, which, while ensuring high precision (accuracy of the predictions made), resulted in a very low recall rate. This low recall rate means that many 'Rejected' proposals were incorrectly classified as 'Approved', leading to a high number of false negatives. This overly cautious approach can be attributed to several factors:

- **Class Imbalance**: The core issue stemmed from the original dataset, which contained far more 'Approved' than 'Rejected' proposals. This imbalance likely caused the model to develop a bias towards predicting proposals as 'Approved'.
- **Influence of External Factors**: Several external factors can influence the decision to reject a proposal, including:
    - **Human Decision-Making.**
    - **Financial Constraints**: Proposals may be rejected due to budget limitations or funding issues, regardless of their inherent quality.
    - **Internal Reasons**: Internal policies or strategic national priorities that are not directly related to the quality of the proposals might also lead to rejections.

● CONCLUSION

This project has successfully demonstrated the application of Text and Data Mining (TDM) techniques to improve the evaluation mechanisms for research proposals in alignment with the Oman Vision 2040. The implementation of a robust classification model, combined with real-time similarity analysis using Multilingual BERT, has yielded several important insights and outcomes:

1. **Alignment with Oman Vision 2040:**
    - By leveraging advanced machine learning and natural language processing techniques, we were able to assess the compatibility of research proposals with the strategic national priorities of Oman Vision 2040. This ensures that the funded research aligns with the long-term developmental goals of the Sultanate, fostering research that contributes to national progress.

2. **Effective Screening Mechanism:**
    - The development and application of a classification model have provided a systematic and automated approach to evaluate competing research proposals. This model, utilizing the RandomForestClassifier and ADASYN for balancing, has proven effective in accurately classifying proposals as 'Approved' or 'Rejected'. This methodology streamlines the evaluation process and enhances the objectivity and consistency of funding decisions.

3. **Improvement of Evaluation Criteria:**
    - Our findings indicate a significant number of proposals classified as 'Approved', with a substantial portion of these actually being 'Rejected'. This highlights the model's conservative nature in avoiding false rejections, ensuring that deserving proposals are not overlooked. However, it also points to the need for refining the evaluation criteria and model parameters to better distinguish high-quality proposals from lower-quality ones.
    - The similarity analysis revealed that some rejected proposals had high alignment scores with Oman Vision 2040. This suggests that while these proposals may align well with strategic priorities, they might have other deficiencies that lead to their rejection. These insights emphasize the importance of a comprehensive evaluation framework that considers both alignment with strategic goals and other critical evaluation criteria.
    - By integrating data-driven insights from the classification model and similarity analysis, we propose adjustments to the existing criteria to ensure they are in line with Oman’s future vision. This will enhance the allocation of research budgets and maximize the impact of funded projects on the local and global community.

4. **Enhancing Funding Opportunities:**

- The implementation of TDM techniques has increased the opportunity to identify and fund research with novel ideas and substantial content. By systematically evaluating proposals, the likelihood of funding innovative projects that address emerging needs and gaps in various sectors is significantly improved.

5. **Comparison with Traditional Evaluation Methods:**
   - Our comparative analysis between traditional paper-based evaluation and modern text mining techniques has demonstrated the effectiveness of automated evaluation methods. The TDM approach not only reduces the time and effort required for evaluation but also provides a more comprehensive and data-driven assessment of research proposals.

Future Implications

The methodologies and findings from this project lay a strong foundation for the continuous improvement of research evaluation processes at Sultan Qaboos University and other research institutions in Oman. The integration of advanced TDM techniques with traditional evaluation frameworks promises to enhance the quality, relevance, and impact of funded research, driving forward the strategic objectives of Oman Vision 2040.

● RECOMMENDATIONS

ii. Adaptation to Changing Criteria:

Given the updates in the RC type criteria in 2024, it’s vital to ensure the model remains relevant and accurate:

- **Dynamic Criteria Updating**: Integrate a system where the model parameters and features can be easily updated to reflect changes in the criteria annually or as needed. This could involve creating a flexible framework within the model that allows for quick adjustments without requiring a complete retraining.
- **Continuous Learning**: Implement a machine learning strategy where the model can learn incrementally as new data comes in, adapting to changes in criteria or new patterns in proposal evaluation.

iii. Development of a User Interface (UI):

- **Interactive Dashboard**: Develop a user-friendly interface that allows users to upload new proposals, view classifications, and receive feedback on the reasons behind each classification. This tool could also allow administrators to adjust the model's parameters in real-time based on evolving needs or feedback.

- **Real-time Classification and Feedback**: Provide a feature where the user can get immediate predictions and explanations, enhancing transparency and trust in the system.

iv. Database for Proposal Tracking:

- **Proposal Tracking System**: Establish a comprehensive database that records all proposals, their classifications, funding status, and any related documents. This system would facilitate easy retrieval and tracking of proposal histories and trends over time.
- **Duplicate and Similarity Checks**: Implement functionality to check new proposals against existing ones in the database to identify potential duplicates or highly similar proposals that have already been funded or rejected.

v. Filtering System Based on Evaluation Criteria:

- **Customizable Filters**: Create a filter system in the UI that allows users to apply different criteria weights or add new criteria as needed. This feature would help in aligning the model more closely with the committee's current evaluation practices and make it more flexible in adapting to new standards.
- **Automated Criteria Checks**: Besides classification, the system could automatically highlight which specific criteria a proposal meets or fails to meet, providing detailed feedback to the evaluators.

vi. Incorporation of External Data Sources:

- **Integration with External Databases**: Link the system with external databases to pull in additional data that can aid in the evaluation process, such as the applicant's previous funding history, publication records, or patent filings.

vii. Long-term Model Assessment and Updates:

- **Regular Model Evaluation**: Set a schedule for regular assessments of the model's performance against new data, ensuring that any drifts in criteria or changes in proposal patterns are promptly addressed.
- **Feedback Loop Mechanism**: Implement a feedback system where users can report discrepancies or suggest improvements, providing data that can be used to further refine the model.

viii. Recommender System:

- **Advanced Recommender System:** Develop a sophisticated recommender system, similar to Applicant Tracking Systems (ATS), that offers full functionality including a UI, dashboard, archive, and robust search capabilities. This system should compare new proposals with past ones to identify patterns and provide insights.
- **Suggestions Based on Trends**: If the system detects that most of the funded proposals in a given year focus on a specific criterion, it should notify users and suggest focusing on other criteria to maintain balance.
- **State-of-the-Art Proposal Suggestions**: Integrate a feature that recommends state-of-the-art projects from around the globe that align with Oman Vision 2040. This will help ensure that the funded proposals are not only innovative but also globally competitive.
- **Proposal Comparison:** Enable the system to compare new proposals with archived ones to highlight similarities, differences, and potential improvements. This feature can provide a historical context and suggest enhancements based on past proposal successes and failures.

● References

البنك الدولي، https://data.albankaldawli.org/indicator/GB.XPD.RSDV.GD.ZS ، تم الدخول بتايخ 16 يناير 2021م.

Anjewierden, A., Kolloffel, B., & Hulshof, C. (2007). Towards educational data mining: Using data mining methods for automated chat analysis to understand and support inquiry learning processes, *International Workshop on Applying Data Mining in e-Learning (ADML 2007),* pp. 23–32.

Dang, S., & Ahmad, P. H. (2014). Text mining: Techniques and its application. *International Journal of Engineering & Technology Innovations*, 1(4), pp. 866-2348.

Geiger, C., Frosio, G., & Bulayenko, O. (2018). The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market-

Legal Aspects. *Centre for International Intellectual Property Studies (CEIPI) Research Paper*, (2018-02).

Hu, W., Dang, A., & Tan, Y. (2019, July). A survey of state-of-the-art short text matching algorithms. In International Conference on Data Mining and Big Data (pp. 211-219). Springer, Singapore.

Johnson, R., Fernholz, O., Fosci, M. (2016) ABDU, *Text and data mining in higher education and public research,* available at : https://adbu.fr/competplug/uploads/2016/12/TDM-in-Public-Research-Revised-15-Dec-16.pdf , (accessed on 2/8/2020).

Ko, Y., & Seo, J. (2000). Automatic text categorization by unsupervised learning. In COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics.

Kobayashi, V.B., Mol, S.T., Berkers, H.A., Kismihók, G. and Den Hartog, D.N., 2018. Text mining in organizational research. Organizational research methods, 21(3), pp.733-765.

Muriru, P.K. and Daewoo, R., “Prediction of the Heat Transfer Characteristics of a Multi-Flame Injector'”, Combustion and Flame, vol. 100, no. 2, pp. 123-135, 2002.

Nordea (2020). The economic context of Oman (Online) Available at: https://www.nordeatrade.com/no/explore-new-market/oman/economical-context, (Accessed on 12 July 2020).

Oman Vision 2040 (Online) Available at : https://www.2040.om/en/, (Accessed on 14 January 2021)

Peters, L., Johnson, M., and Davidson, K., “A Novel Approach to Four-Bar Synthesis'”, 10th ASME Design Automation Conference, pp. 234-250, Pittsburgh, PA, 2001.

San Tay, P., & Sik, C. P. (2016). Data mining and copyright: A bittersweet technology gift for copyright owners and the Malaysian public?. *Computer Law & Security Review*, 32(6), pp. 898-906.

Swanson Inc., “Online Users Manual for ANSYS 5.0'', http://www.ansys.com/ manual, viewed on March 1999.

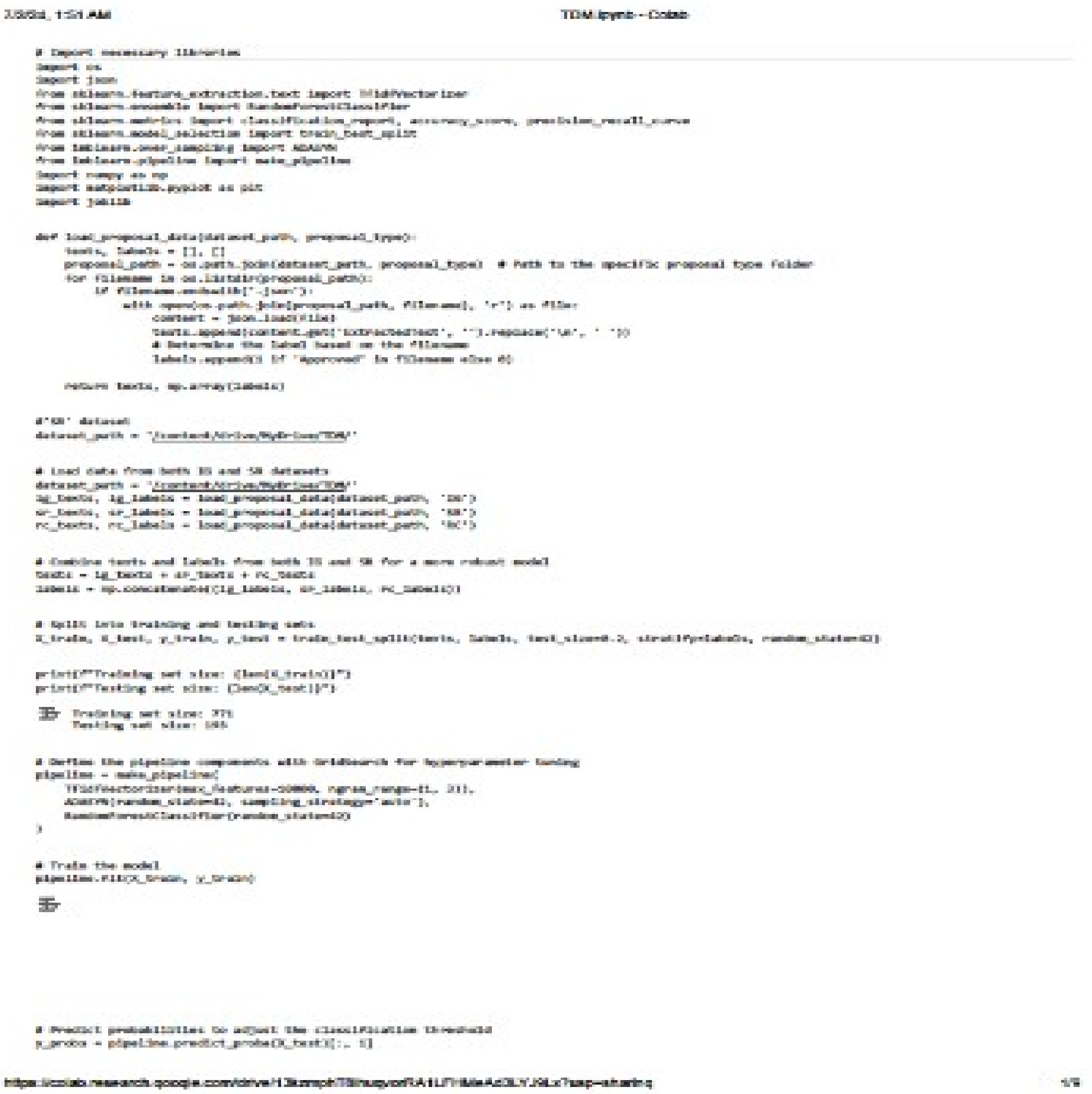
Truyens, M., & Van Eecke, P. (2014). Legal aspects of text mining. *Computer law & security review*, 30(2), pp. 153-170.

Wen-Cheng, C., “Electric Bicycle'”, US Patent no. 5,368,122, November 29, 1994.

Zacharia, M. and Daudi, P.K., The Effect of Multi-materials on Conventional Finite Element Formulations, New York: Wiley and Sons, 2001.

5. APPENDIX B

Our Notebook



7/3/24, 1:51 AM TDM.ipynb - Colab

```
# Import necessary libraries
import os
import json
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, accuracy_score, precision_recall_curve
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import ADASYN
from imblearn.pipeline import make_pipeline
import numpy as np
import matplotlib.pyplot as plt
import joblib

def load_proposal_data(dataset_path, proposal_type):
    texts, labels = [], []
    proposal_path = os.path.join(dataset_path, proposal_type)  # Path to the specific proposal type folder
    for filename in os.listdir(proposal_path):
        if filename.endswith('.json'):
            with open(os.path.join(proposal_path, filename), 'r') as file:
                content = json.load(file)
                texts.append(content.get('ExtractedText', '').replace('\n', ' '))
                # Determine the label based on the filename
                labels.append(1 if 'Approved' in filename else 0)

    return texts, np.array(labels)

#'SR' dataset
dataset_path = '/content/drive/MyDrive/TDM/'

# Load data from both IG and SR datasets
dataset_path = '/content/drive/MyDrive/TDM/'
ig_texts, ig_labels = load_proposal_data(dataset_path, 'IG')
sr_texts, sr_labels = load_proposal_data(dataset_path, 'SR')
rc_texts, rc_labels = load_proposal_data(dataset_path, 'RC')

# Combine texts and labels from both IG and SR for a more robust model
texts = ig_texts + sr_texts + rc_texts
labels = np.concatenate((ig_labels, sr_labels, rc_labels))

# Split into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(texts, labels, test_size=0.2, stratify=labels, random_state=42)

print(f"Training set size: {len(X_train)}")
print(f"Testing set size: {len(X_test)}")
```

```
Training set size: 771
Testing set size: 193
```

```
# Define the pipeline components with GridSearch for hyperparameter tuning
pipeline = make_pipeline(
    TfidfVectorizer(max_features=10000, ngram_range=(1, 2)),
    ADASYN(random_state=42, sampling_strategy='auto'),
    RandomForestClassifier(random_state=42)
)

# Train the model
pipeline.fit(X_train, y_train)
```

```
# Predict probabilities to adjust the classification threshold
y_probs = pipeline.predict_proba(X_test)[:, 1]
```

https://colab.research.google.com/drive/[illegible] 1/9