RUTGERS

TEXTUAL ANALYSIS IN FINANCE

FINAL PROJECT REPORT

Textual Analysis of Mutual Fund Prospectus Report & Prediction of Expense Ratio

NAME: Mohammed Nihil Puthiya Kottal
ID: 227001556

NetID: mnp126

ABSTRACT

This project analyses Scheme Information Documents (SIDs) and Statement of Additional Information (SSDs) of 60 mutual funds, sourced directly from the respective fund house websites, to extract and interpret meaningful insights. The methodology integrates advanced natural language processing (NLP) techniques and statistical modelling. Semantic analysis was performed using the LM dictionary to evaluate the linguistic attributes of the documents, complemented by the FOG index to assess readability and comprehension complexity. The analysis revealed that these documents contain high-complexity content. The ChatGPT API was leveraged to generate concise summaries of the extracted information, enhancing accessibility for non-expert users. Furthermore, cosine similarity was applied to identify mutual funds with similar textual characteristics, enabling comparative analysis. A regression model was subsequently implemented to predict the expense ratio of mutual funds based on textual features from the documents. This comprehensive approach combines NLP, text summarization, and predictive modelling to provide a novel perspective on mutual fund analysis

INTRODUCTION

Mutual funds are a cornerstone of investment strategies, offering a structured approach to diversifying portfolios and managing financial risks. However, the complexity and length of Scheme Information Documents (SIDs) and Statement of Additional Information (SSDs) pose significant challenges for investors. These documents are rich in detail, outlining the fund’s objectives, strategies, and associated costs, yet their dense language often makes them inaccessible to the average investor. This inaccessibility can hinder informed decision-making and create barriers for those seeking to maximize the benefits of mutual fund investments.

The importance of analyzing these documents lies in their role as essential disclosures mandated by regulatory bodies. Understanding the content of SIDs and SSDs can empower investors by clarifying fund-specific details such as expense ratios, risk profiles, and investment strategies. Despite this, existing tools for analyzing mutual fund documents are limited in extracting and summarising meaningful information, leaving a gap in the market for advanced analytical solutions.

Recent advances in natural language processing (NLP) and machine learning have opened new possibilities for extracting and interpreting textual data. While several studies have explored the application of NLP to financial documents, none have specifically focused on the detailed analysis of SIDs and SSDs to uncover insights like document readability, fund similarity, and cost predictability. This project addresses this research gap by combining cutting-edge NLP techniques and predictive modelling to analyse mutual fund documents comprehensively.

The study employs semantic analysis using the LM dictionary to evaluate the linguistic attributes of the documents alongside the FOG index to measure readability and complexity. ChatGPT API calls generate concise summaries, making dense information more accessible. Furthermore, cosine similarity analysis identifies mutual funds with similar textual characteristics, enabling easier investor comparisons. Finally, a regression model predicts expense ratios based on the textual features of these documents, offering valuable insights into cost structures.

By leveraging these advanced methodologies, this project aims to enhance the understanding of mutual fund documentation and seeks to empower investors with tools to make more informed decisions.

LITERATURE REVIEW

Over the years, advancements in natural language processing (NLP), similarity measures, and regression modelling have paved the way for innovative approaches to analyzing and interpreting textual data across various domains. This review consolidates insights from existing research that underpin the methodologies employed in this project, including NLP techniques, cosine similarity, and regression analysis.

Semantic analysis and text processing have been extensively studied to extract meaningful insights from unstructured data. Jurafsky and Martin [1] provide a foundational framework for understanding NLP techniques, emphasizing the role of preprocessing, tokenization, and feature extraction in text analysis. Loughran and McDonald [2] applied semantic dictionaries to financial texts, highlighting the importance of domain-specific lexicons, such as the LM dictionary, for uncovering linguistic patterns in financial documents. Similarly, Mikolov et al. [3] introduced word embeddings that capture semantic relationships, demonstrating the potential of distributed representations in improving the interpretability of textual data.

The FOG index, a widely used readability metric, has also gained prominence in assessing document complexity. Gunning [4] introduced the metric to evaluate the accessibility of written content, which has since been applied to various domains, including financial documents, to quantify their comprehensibility.

Cosine similarity remains a cornerstone in text analysis for quantifying the similarity between textual documents. Salton and McGill [5] pioneered its use in information retrieval, demonstrating its effectiveness in comparing text based on vector representations. Huang [6] expanded on its applications by exploring its role in clustering and classification tasks, particularly for unstructured text data. These studies underscore the importance of cosine similarity in identifying relationships between documents, as employed in this project to find similar mutual funds based on their Scheme Information Documents (SIDs) and Statement of Additional Information (SSDs).

Regression analysis has long been a fundamental tool for predictive modelling. Hastie et al. [7] outlined the applications of regression in identifying relationships between independent and dependent variables, providing a comprehensive understanding of linear and nonlinear regression techniques. Montgomery et al. [8] emphasized the importance of feature selection and dimensionality reduction in regression modelling to improve prediction accuracy and interpretability. Recent advances in regression analysis have incorporated textual features as predictors, bridging the gap between unstructured data and quantitative modelling. Pedregosa et al. [9] demonstrated the versatility of regression models implemented in Scikit-learn, showcasing their adaptability for tasks involving complex, multi-dimensional datasets.

METHODOLOGY

This study employs a combination of natural language processing (NLP), cosine similarity, and regression modelling to analyze Scheme Information Documents (SIDs) and Statement of Additional Information (SSDs) of mutual funds.

1. **Natural Language Processing (NLP)**
   - **Semantic Analysis:** The Loughran-McDonald (LM) dictionary was used to evaluate linguistic features specific to financial texts.
   - **FOG Index:** Readability was assessed using the FOG index, highlighting the complexity of the documents.
2. **Document Summarization:** ChatGPT API generated concise summaries, distilling critical information like investment strategies and expense ratios for easier interpretation.
3. **Cosine Similarity:** Cosine similarity was applied to compare mutual funds by converting document content into vector representations, enabling the identification of similar funds based on textual features.
4. **Regression Modelling:** Textual features were used as predictors in a linear regression model to forecast expense ratios, establishing relationships between document attributes and fund costs.

Data Collection

The dataset utilized in this study was sourced directly from the official websites of three prominent fund houses: IDBI Mutual Fund, ICICI Prudential Mutual Fund, and TATA Mutual Fund. The Scheme Information Documents (SIDs) and Statement of Additional Information (SSDs) of 60 mutual funds were collected for analysis. These documents provided detailed information on fund objectives, strategies, risks, and expense structures, forming the basis for this study's textual and predictive analyses.

- ***IDBI Mutual Fund***: *https://www.idbimutual.co.in/Downloads/SID#*
- ***ICICI Mutual Fund:*** *https://www.archive.icicipruamc.com/downloads/ssd*
- ***TATA Mutual Fund:*** *https://www.tatamutualfund.com/schemes-related/scheme-summary*

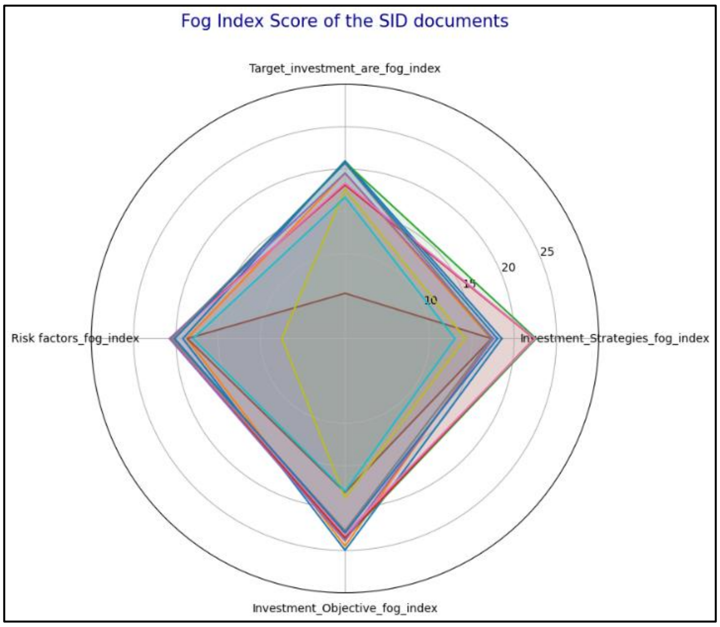
Readability Assessment

The fog index was calculated to gauge the complexity of the document:

Sentence and Word Analysis: Calculates the average length of sentences and identifies complex words (words with three or more syllables).

Complexity Score: Computes the percentage of complex words and combines it with the average sentence length to determine the Fog Index using a specific formula.

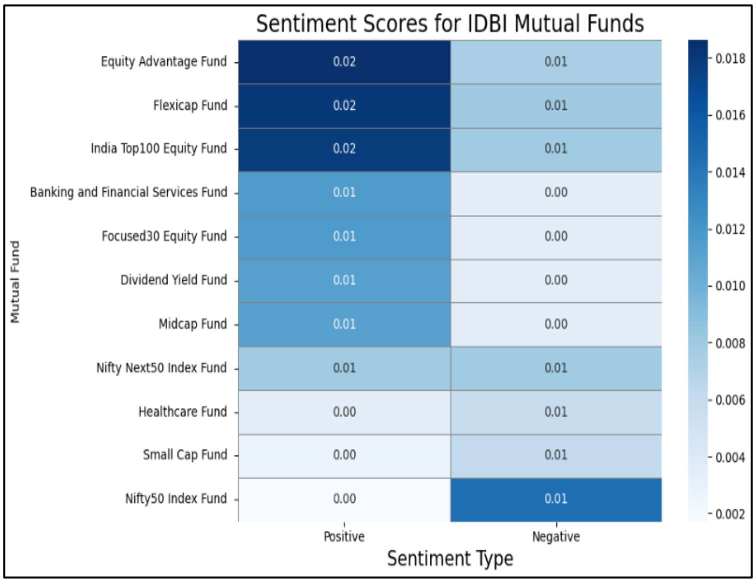
.



Sentiment Analysis

The Loughran-McDonald sentiment dictionary identifies and counts words associated with specific emotional sentiments. The analysis captures the most frequent sentiment-laden words, offering insights into the predominant emotional tones conveyed in the documented sections.

- A positive tone highlights growth potential and investor benefits, enhancing investor confidence and trust
- The consistently low negative tone across all funds reduces perceived risks, ensuring the documents remain factual and professional, which helps maintain investor assurance.

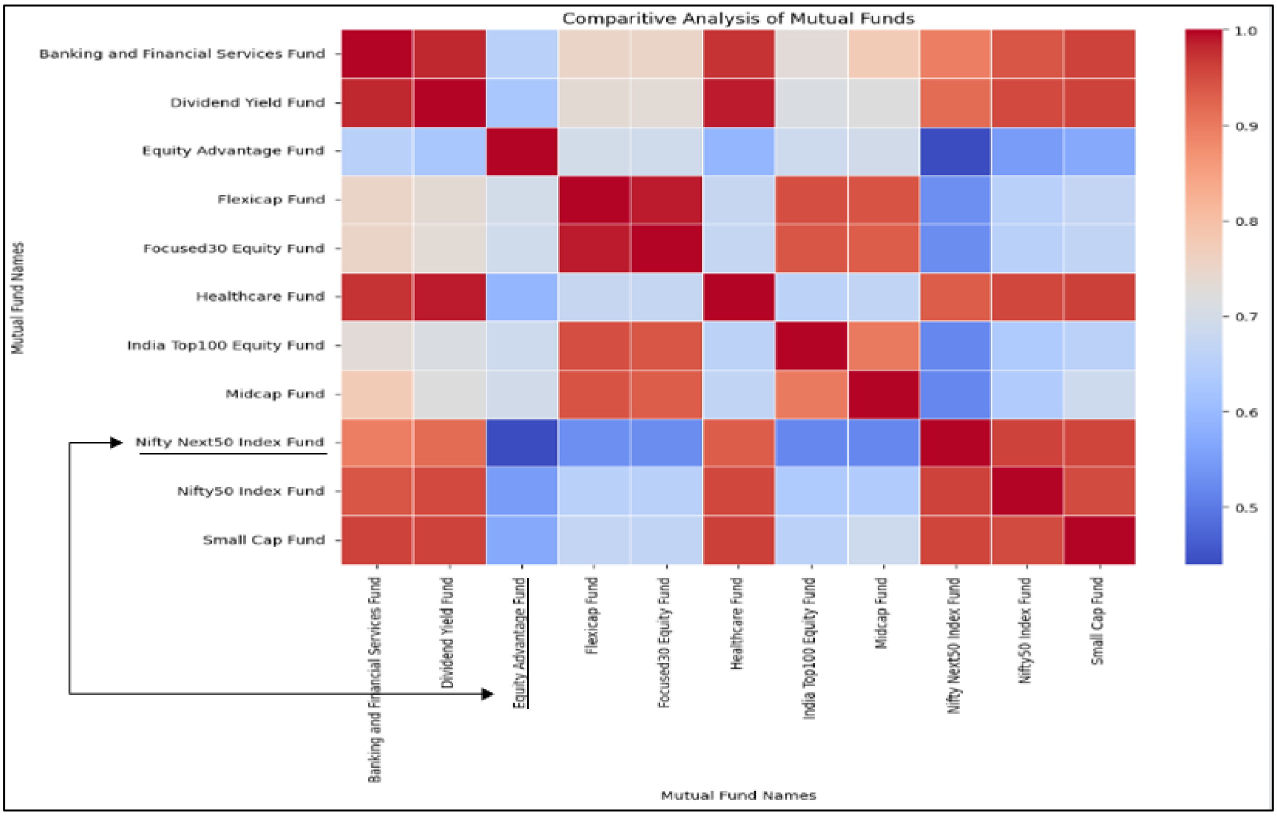


From the figure, Investors and analysts can use this sentiment analysis better to understand the tone and implications of the fund descriptions, aiding in more informed decision-making by highlighting how fund managers communicate about their funds.

Comparative Analysis of Mutual Funds

The analysis aimed to compare and find similarities among mutual funds based on several key aspects: investment objective, investment strategy, risk factors, and target investment area. To achieve this, the text from these specific columns was combined and analyzed using cosine similarity, a method commonly used in text analysis to measure similarity between documents.

Cosine similarity works by converting text documents into vectors of terms using methods like TF-IDF (Term Frequency-Inverse Document Frequency). Each document is represented as a vector in a multidimensional space, where each dimension corresponds to a unique term in the overall corpus of documents. Cosine similarity then measures the cosine of the angle between these two vectors. An angle of 0 degrees, or a cosine similarity score of 1, indicates that the documents are identical in terms of term composition, while a 90-degree angle, or a score of 0, shows that the documents share no terms. This approach allows for an efficient and effective comparison of documents, pinpointing mutual funds with similar characteristics based on their textual content.



*Based on the generated cosine similarity score in the above figure, the highlighted mutual funds '**Nifty Next50 Index Fund' and 'Equity Advantage Fund'** completely differ regarding their Investment Objectives, Investment Strategy, Risk Factors and target Investment Area.*

In-depth Analysis of Mutual Fund

After employing cosine similarity to assess the textual similarities across various mutual funds, the next logical step was to delve deeper into the content of an individual mutual fund to understand its specific focus and thematic elements. This is crucial because, while cosine similarity provides a valuable quantitative measure of how closely the content of one mutual fund resembles another, it does not reveal the qualitative aspects of what the mutual fund is talking about.

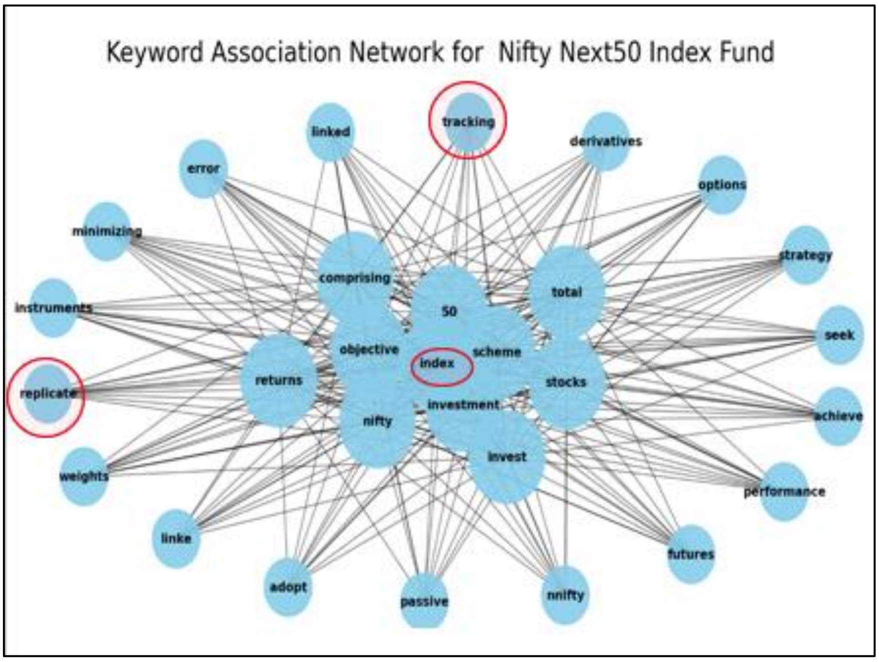
Cosine similarity essentially offers a broad comparison, allowing us to see which funds are textually similar and which are not. Still, it leaves out the details of the underlying content—what themes, strategies, and specifics are emphasized in the mutual fund's documents.

To bridge this gap, the keyword association network was introduced as a method to visually and analytically parse the content of an individual mutual fund, in this case, the Nifty Next50 Index Fund. This approach involves creating a network graph that maps out the relationships between keywords used in the fund's documentation. This graphical representation helps identify:

- **Key Themes**: By analyzing the central keywords and their associations, we can discern the main themes around which the fund's content is structured.

- **Strategic Focus**: The relationships between words like "index", "investment", "tracking", and "replicant" provide insights into the fund's investment strategies, such as whether it aims to replicate the performance of an index.

- **Operational Details**: Terms connected to operational tactics, like "minimizing error" or "adjusting weights", can reveal how the fund manages its portfolio and aims to achieve its objectives.



This detailed analysis complements the findings from the cosine similarity and adds a layer of depth by highlighting what the mutual fund focuses on and how it communicates its strategies and objectives to investors. Such an analysis is instrumental for stakeholders who need a deeper understanding of a fund’s content beyond knowing its similarity to other funds, aiding in better-informed investment decisions based on specific fund characteristics and strategies.

Prediction of Expense Ratio

The project aimed to develop a predictive model that could forecast the expense ratios of mutual funds based on their textual features derived from Scheme Information Documents (SIDs) and Scheme Summary Documents (SSDs). This approach is innovative as it links document characteristics, such as complexity and thematic content, directly to cost metrics, thereby providing insights into how the textual content of a fund's documentation might influence its operational expenses.

In the project, a Ridge Regression model was employed. Ridge Regression, also known as Tikhonov regularization, estimates the coefficients of multiple-regression models in scenarios where independent variables are highly correlated. This technique is particularly useful in preventing overfitting and improving the model's generalization by introducing a regularization term.

Feature Extraction:

- **TF-IDF Vectorization:** Textual features were extracted using TF-IDF on the top 100 features ranging from bigrams to four-grams, highlighting key terms relevant to mutual fund descriptions

- **Numeric Features:** Additional numeric features like ‘Risk_Level_Numeric’ and ‘fund_type_numeric’ were incorporated to capture the risk profile and type of the mutual funds, enhancing the model’s predictive capabilities.

Model Construction:

- **Initial Setup:** The Ridge Regression model was initially configured with an alpha parameter of 1.0. This parameter helps control the model's complexity and prevents overfitting by penalizing larger coefficients.

- **Model Evaluation:** The model’s performance was assessed using R-Squared and RMSE metrics to determine its accuracy and the average magnitude of prediction errors.

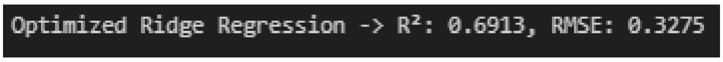
Hyperparameter Tuning:

- **Grid Search:** A grid search was conducted to optimize the alpha parameter, ensuring the best model performance by systematically exploring various alpha values to minimize prediction errors.

Model Refinement and Validation:

- **Optimized Model:** After finding the optimal alpha value, the model was retrained to fine-tune its predictions, aiming for improved accuracy
- **Residual Analysis:** A thorough analysis of residuals was performed to ensure no significant bias or pattern, verifying that the model predictions were reliable and unbiased

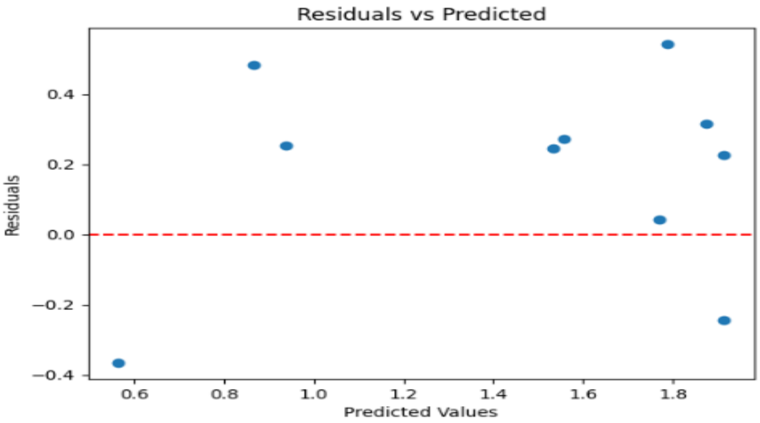
Model Result



```
Optimized Ridge Regression -> R²: 0.6913, RMSE: 0.3275
```

An R^2 value of 0.69 metric indicated a good proportion of variance in the mutual fund expense ratios that the model successfully explained.

Residual Analysis: The absence of patterns in the residual analysis points to the robustness of the model. This means the model is reliable for practical use, providing stakeholders with trustworthy predictions of expense ratios based on the textual content of mutual fund documents.



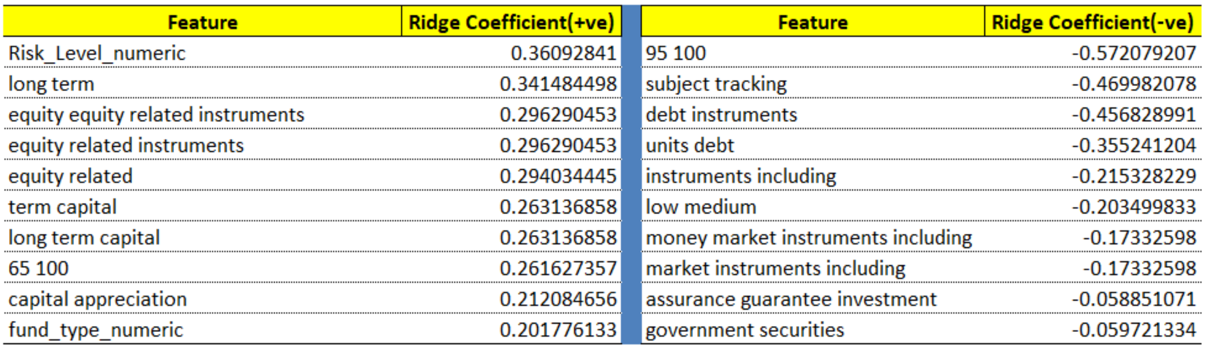
Positive Coefficients (Increase in Expense Ratio):

- **Risk Level Numeric (0.3609)**: Higher risk levels in mutual funds are positively correlated with higher expense ratios, indicating that funds positioned as higher risk might involve more complex management strategies that increase operational costs.
- **Long Term (0.3415)**: References to "long term" investments are associated with higher expense ratios, possibly due to the extended time horizon requiring ongoing fund management and oversight

Negative Coefficients (Decrease in Expense Ratio):

- **95 100 (-0.5727):** This feature, potentially a category or a specific attribute of funds, shows a strong negative association with expense ratios, indicating that this feature characterizes funds with lower operational costs.

- **Debt Instruments (-0.4568)**: Funds emphasizing debt instruments tend to have lower expense ratios, consistent with the generally lower fees associated with managing bond portfolios compared to equities.



Feature	Ridge Coefficient(+ve)	Feature	Ridge Coefficient(-ve)
Risk_Level_numeric	0.36092841	95 100	-0.572079207
long term	0.341484498	subject tracking	-0.469982078
equity equity related instruments	0.296290453	debt instruments	-0.456828991
equity related instruments	0.296290453	units debt	-0.355241204
equity related	0.294034445	instruments including	-0.215328229
term capital	0.263136858	low medium	-0.203499833
long term capital	0.263136858	money market instruments including	-0.17332598
65 100	0.261627357	market instruments including	-0.17332598
capital appreciation	0.212084656	assurance guarantee investment	-0.058851071
fund_type_numeric	0.201776133	government securities	-0.059721334

The coefficients derived from the Ridge Regression model provide a detailed quantification of how specific textual features and fund characteristics influence the operational costs of mutual funds, as measured by expense ratios. This methodology enables a robust approach to developing and evaluating predictive models for estimating mutual fund expense ratios, emphasizing the practical applicability of predictions in financial management and investment strategy formulation.

By focusing on both positive and negative influences and understanding the weight each feature holds in the model, fund managers and stakeholders can make more informed decisions regarding fund documentation, investment strategies, and cost management. This analysis is crucial for tailoring fund offerings to meet specific financial goals and ensuring that the fund's operational costs are transparently communicated to potential investors.

Implementation

- **Software and Tools**: The analysis was performed using Python & R programming language, which is well-suited for statistical modelling and machine learning applications. Relevant packages such as PyPDF2 to extract text from PDF files, Sklearn & nltk for regression & textual analysis, openai for ChatGPT summary, matplotlib, network and seaborn were used for visualization.

CONCLUSION

In conclusion, this project successfully integrated a range of sophisticated data analysis techniques, including natural language processing (NLP), cosine similarity, and Ridge Regression, to provide a comprehensive overview of mutual fund documents and predict expense ratios based on their textual content. Here's how each component contributed to the project's objectives:

- **Textual Analysis and Simplification:** By employing the Fog Index to assess readability, the project helped simplify complex mutual fund documents, making them more accessible to average investors. This was crucial in enhancing transparency and comprehensibility, addressing the often opaque nature of financial documents.

- **Keyword and Sentiment Analysis:** The analysis of key terms and sentiments using the Loughran-McDonald dictionary provided insights into the thematic emphasis and emotional tone of the documents. This helped identify the growth potential and risk perceptions portrayed in the funds, enhancing investor understanding and confidence.

- **Cosine Similarity for Comparative Analysis:** Using cosine similarity, the project effectively grouped mutual funds with similar characteristics, facilitating easier comparison for investors. This comparative analysis helped highlight similarities and differences between funds, aiding in informed decision-making

.

- **Predictive Modelling:** The Ridge Regression model predicted expense ratios by linking textual features from the fund documents to cost metrics. This predictive approach provided fund managers and investors with a data-driven tool to anticipate fund costs, optimize expenses, and benchmark against industry standards.

- **Practical Application and Impact:** The methodologies employed in this project not only advanced the analytical capabilities of financial document analysis but also had a direct impact on stakeholder decision-making. By making mutual fund documents more understandable and providing a predictive outlook on expense ratios, the project supported better investment strategies and regulatory compliance.

- **Broader Implications:** Beyond immediate project outcomes, the insights and techniques developed contribute to broader applications in financial analytics. They offer a framework for other financial institutions to adopt similar strategies, potentially leading to industry-wide improvements in fund documentation clarity and investment transparency.

In summary, this project has set a new standard in the analysis of mutual fund documents by seamlessly combining NLP, statistical modeling, and machine learning. It has enhanced the accessibility, comparability, and understanding of mutual fund offerings, thereby empowering investors and fund managers with critical insights and tools for better financial governance. The success of this project underscores the potential for data science to revolutionize financial document analysis and investment strategy development.

REFERENCES

1. Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (3rd Edition). Pearson.
2. Loughran, T., & McDonald, B. (2011). "When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *Journal of Finance*, 66(1), 35–65.
3. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). "Efficient Estimation of Word Representations in Vector Space."
4. Gunning, R. (1952). *The Technique of Clear Writing*. McGraw-Hill.
5. Salton, G., & McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill.
6. Huang, A. (2008). "Similarity Measures for Text Document Clustering." *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC)*, 49–56.
7. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd Edition). Springer.
8. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th Edition). Wiley..
9. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, 12, 2825–2830.
10. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.