Extracting Key Insights from Financial Text: A Summarization Approach

Parisha Desai
Computer Science Department
Purdue University Fort Wayne
2101E Coliseum Blvd,
Fort Wayne, IN 46805
desap01@pfw.edu

Manvitha Chowdari Gottipati
Computer Science Department
Purdue University Fort Wayne
2101E Coliseum Blvd,
Fort Wayne, IN 46805
gottm01@pfw.edu

Abstract

The financial industry generates vast amounts of textual data vital for decision-making, but efficiently processing this data is challenging due to information overload. This study examines advanced natural language processing (NLP) techniques for automated summarization of financial text, utilizing domain-specific datasets and hybrid methods. We explore extractive models like TextRank and state-of-the-art abstractive models such as T5, BART, and Pegasus, focusing on preprocessing, fine-tuning, and hyperparameter optimization to preserve numerical accuracy and domain-specific details. Performance evaluation using metrics such as ROUGE, BERTScore, and METEOR reveals that Pegasus and T5 excel in generating coherent summaries, though challenges like low BLEU scores and hybrid model limitations remain. This work advances financial summarization techniques, providing scalable solutions for extracting critical insights and supporting informed decision-making.

1 Introduction

The financial sector produces extensive textual data in the form of reports, market statements, and news. Processing this immense volume of information is critical for informed decision-making but presents challenges due to the effort required. Our motivation stems from the critical need for timely and accurate insights in the financial domain, which are essential for risk assessment and maximizing returns. This study addresses these challenges by developing and refining summarization models that can effectively process financial terminology and context. Automating the summarization process with NLP-based models saves time while ensuring precision. Unlike other summarization methods, financial narrative summarization uniquely requires the preservation of domain-specific details and numerical accuracy. This work builds on prior research by integrating transformer models and domain-specific optimizations, aiming to bridge gaps in coherence, relevance, and domain adaptation.

2 Related Work

Financial text summarization has garnered significant attention, with methods ranging from traditional statistical approaches to advanced machine learning models. TextRank, a widely used graph-based extractive summarization algorithm, excels in identifying key sentences by ranking their graph connectivity and similarity (Mihalcea, 2004). (Zhang et al., 2024) surveyed summarization techniques in the financial domain, highlighting the evolution from basic statistical methods to large language models tailored for financial tasks.

Recent efforts have leveraged hybrid models, combining extractive techniques with the generative capabilities of T5, BART, and Pegasus. These approaches, as demonstrated by (Liu et al., 2023), have yielded promising results, producing more coherent and concise summaries. Hybrid models are particularly promising in financial contexts, as they combine the semantic coherence of abstractive techniques with the precision of extractive approaches, addressing challenges like redundancy and relevance inherent in financial texts.

3 Methodology

Our methodology consisted of several key steps designed to effectively summarize financial text, including dataset analysis, preprocessing, the identification of summarization methodologies and models, model implementation, and performance evaluation. Each step was carefully crafted to address the challenges of summarizing domain-specific financial narratives.

3.1 Dataset Analysis

The dataset[1] used in this research originates from the Financial Narrative Summarization (FNS) task presented at FNP 2020. It comprises annual reports from publicly traded companies alongside their corresponding human-written summaries. These reports, often exceeding 40,000 words per record, contain unstructured text highlighting financial performance, operational activities, and critical business insights. The gold-standard summaries, averaging 1,000 words per record, are structured to capture essential information.

The dataset is divided into three subsets: 2,000 records in the training set, and 250 records each in the validation and test sets.

3.2 Preprocessing Techniques

Preprocessing was a key step in ensuring the dataset's integrity and relevance. Initial preprocessing involved tasks such as removing symbols and cleaning records with missing data. Stop word removal and lemmatization further enhanced data quality by reducing redundancy and standardizing word forms. These steps were critical in preparing the dataset for effective model training and evaluation. The original text case, numerical data, and short words (fewer than three characters) were preserved to retain critical contextual significance,which is essential in financial analysis.

3.3 Identifying Summarization Methodologies and Models

Two primary summarization methodologies, extractive and abstractive, were explored, with specific models selected for each category. Extractive summarization identifies and selects the most important sentences from the source text to create concise summaries. Within this category, the TextRank model, a graph-based and unsupervised summarization approach, was employed. TextRank represents sentences as nodes in a graph, with edges indicating similarity, and generates summaries by ranking sentences based on importance. Abstractive summarization, on the other hand, generates summaries by rephrasing and reorganizing the source text using advanced natural language processing (NLP) techniques. Models utilized in this category included T5, a pretrained text-to-text transformer with an encoder-decoder architecture; BART, which combines a bidirectional encoder (like BERT) with an autoregressive decoder (like GPT); and Pegasus, a model specifically designed for summarization that uses Gap Sentence Generation (GSG) to predict masked sentences and create content-rich summaries.

The end-to-end training approach proposed by La Quatra and Cagliero (2020) serves as a comparative benchmark for this study, particularly in the implementation of abstractive models. Expanding on their insights, this project adopts a hybrid framework to enhance coherence and fluency in financial text summaries. Specifically, a two-stage pipeline combining Pegasus and T5 was implemented, where Pegasus generated the initial summaries and T5 refined them to improve fluency and coherence. To optimize performance, the best hyperparameter settings for each independent model were integrated into the hybrid framework.

3.4 Model Implementation

The model implementation [2] followed a structured approach to ensure optimal performance. A summarization function was designed using structured prompts, such as 'Summarize: annual report,' to guide the models in generating summaries. The integration of tokenizers, encoders, and decoders was tailored to the specifications of each selected model. Inspired by the methodologies employed by Abdaljalil and Bouamor (2021), this research integrates domain-specific optimizations with state-of-the-art summarization models to address the complexities of financial text. Their emphasis on preserving domain-specific details has been a pivotal consideration in model fine-tuning and evaluation.

Fine-tuning was performed on all models, except TextRank (an unsupervised model), using the training dataset to capture domain-specific patterns and nuances present in financial texts. In addition, parameter optimization was performed by experimenting with variables such as minimum summary length and beam size. These adjustments were aimed at balancing the quality of summarization and computational efficiency, ensuring robust model performance.

[1] Dataset:https://huggingface.co/datasets/ragha92/FNS_Summarization/tree/main

[2] Source code is available at: https://github.com/gottm01/Text_Summarization_Financial_Text/tree/main/Final_Update

3.5 Performance Evaluation

Model performance was evaluated using various metrics to capture various aspects of summarization quality. ROUGE, which measures n-gram overlap between generated and reference summaries, focused on content coverage and recall through ROUGE-1 (unigrams), ROUGE-2 (bigrams), and ROUGE-L (longest common subsequence). BLEU assessed precision by comparing n-gram sequences in the generated and reference summaries, emphasizing syntactic alignment while penalizing structural differences. BERTScore leveraged pretrained embeddings, such as RoBERTa, to evaluate semantic similarity, effectively capturing deeper contextual alignment essential for accurate meaning representation. METEOR accounted for exact matches, synonyms, paraphrases, and word order, enabling the recognition of nuanced expressions and variations common in abstractive summarization. Optimal value for all these metrics are closer to 1. These metrics provided a robust evaluation framework, balancing content accuracy, semantic coherence, and contextual accuracy. Models were rigorously tested on validation and test datasets to assess generalization and identify areas for improvement.

4 Our Experimentation

The experimentation phase involved iterative refinements to preprocessing, fine-tuning, and hyperparameter optimization. Initial preprocessing steps, such as removing whitespace and trailing characters, evolved to include advanced techniques like removing special characters and applying lemmatization. These enhancements demonstrated slight improvements in model performance, underscoring the importance of clean and structured input data.

Fine-tuning played a pivotal role in adapting the models to financial text. Parameters such as minimum token length and length penalty were systematically adjusted to improve summary quality. For example, increasing the minimum token length criterion ensured that generated summaries retained essential details, while reducing the length penalty improved content fluency and helped our models generate summaries that closely matched the gold summary.

Hyperparameter optimization was conducted using Optuna with 10 trials, systematically tuning parameters such as token length, learning rate, length penalty, number of beams, and top rank. For example, increasing the number of beams enhanced summary accuracy by enabling the model to evaluate a broader range of sentences, however it lead to increased computational costs in terms of time and resource consumed in a exceution. Finally, the evaluation framework, initially limited to ROUGE and BLEU, was expanded to include METEOR and BERTScore. These additional metrics provided a more comprehensive assessment of model performance, capturing semantic and paraphrasing aspects that were not effectively measured by ROUGE and BLEU alone.

Model Name	Min Length	Max Length	Learning Rate	Length Penalty	No. of Beams	Top Rank
BART	150	576	2.32e-05	1.91	5	N/A
Pegasus	100	512	1.313-05	1.76	9	N/A
T5	50	896	2.49e-04	1.56	7	N/A
TextRank	N/A	N/A	N/A	N/A	N/A	2

Table 1: Optimal-Hyperparameters

Table:1 illustrates the optimal hyperparameters identified for each model

5 Results and Observations

Tables 2 and 3, along with Figures 1 and 2, provide a detailed comparison of performance metrics for all evaluated models on the validation and test datasets, respectively.

It can be observed that T5 and Pegasus stood out as the top-performing models across both the validation and test datasets, maintaining consistent ROUGE and BERTScore values. This consistency highlights their robustness in summarizing complex financial narratives. T5's encoder-decoder architecture and Pegasus's domain-specific pretraining likely contributed to their strong performance. However, the hybrid model combining Pegasus and T5 did not achieve the expected performance gains. This outcome suggests that the current ensemble strategy requires further refinement. Potential improvements could include experimenting with alternative ways to combine the outputs of the two models or exploring advanced fine-tuning techniques to better align their strengths.

TextRank, despite being an unsupervised extractive summarization model, achieved the highest BERTScore, demonstrating its ability to extract sentences with high semantic similarity to the gold summaries. This makes it a reliable baseline for the project, particularly in highlighting key semantic elements of the source text. However, its performance on other metrics, such as ROUGE and BLEU, was comparatively weaker due to its inherent limitations in producing paraphrased or restructured outputs. BLEU scores, on the other hand, were notably low across all models. This can be attributed to BLEU's strict reliance on exact word overlap, which inherently penalizes paraphrasing—a defining characteristic of abstractive summarization models like T5 and Pegasus. BART, another abstractive summarization model, performed poorly across all metrics. This underperformance may stem from its pretraining on a general corpus, limiting its capability to effectively understand and summarize domain-specific financial narratives. Nonetheless, the alignment of results across validation and test datasets for all models indicated strong generalization capabilities and minimized overfitting. This consistency underscores the balanced and high-quality annotation of the dataset, which provided a robust foundation for model training and evaluation.

Model Name	**ROUGE-1 Score**	**ROUGE-2 Score**	**ROUGE-L Score**	**BLEU Score**	**METEOR Score**	**BERT Score**
T5	**0.2333**	**0.1372**	0.1668	**0.0898**	**0.1961**	0.7922
BART	0.1739	0.0134	0.1047	0.0015	0.0596	0.7405
Pegasus	0.2334	0.1312	**0.1738**	0.0792	0.1853	0.7959
Hybrid	0.1236	0.0597	0.0885	0.0220	0.0737	0.7823
TextRank	0.0716	0.0681	0.0690	0.0375	0.1310	**0.8105**

Table 2: Validation Dataset Results

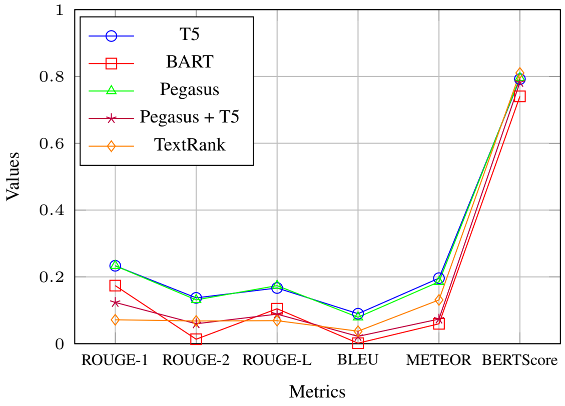

Figure 1: Model performance on Validation Dataset

Model Name	**ROUGE-1 Score**	**ROUGE-2 Score**	**ROUGE-L Score**	**BLEU Score**	**METEOR Score**	**BERT Score**
T5	**0.2130**	**0.1096**	0.1393	**0.0623**	**0.1698**	0.7880
BART	0.1695	0.0127	0.1027	0.0015	0.0578	0.7382
Pegasus	0.2121	0.1095	**0.1464**	0.0609	0.1585	0.7924
Hybrid	0.1018	0.0444	0.0718	0.0123	0.0594	0.7770
TextRank	0.0762	0.0704	0.0728	0.0399	0.1300	**0.8030**

Table 3: Test Dataset Results

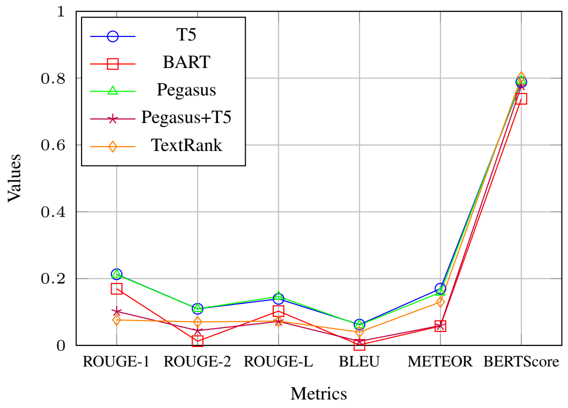

Figure 2: Model performance on Test Dataset

6 Additional Analysis

Given T5 and Pegasus's strong performance, future efforts could focus on domain-specific pretraining or advanced fine-tuning methods to enhance their capacity for summarizing financial reports.The interplay between beam size and computational cost was evident in experiments. Smaller beam sizes improved ROUGE scores but limited sentence diversity, aligning with the goal of producing summaries that closely match the gold standard. Low BLEU scores across all models emphasize the limitations of relying solely on word-overlap-based metrics for abstractive summarization tasks. This finding suggests that semantic-focused metrics like BERTScore and METEOR are more reliable for evaluating financial text summaries as models would not be able to generate text capturing the variability inherent in human interpretation-based Gold summaries.

The underperformance of the Pegasus + T5 model suggests that further exploration of hybrid approaches, such as cascading models with feedback loops or weighted ensembling, could be beneficial. Wherein cascading models pass the output of one model to another for refinement, with feedback loops to adjust and improve their outputs dynamically. Weighted ensembling blends outputs from multiple models, assigning higher weights to models based on their strengths, such as accuracy or fluency, to achieve balanced and effective summaries. BERTScore's high performance across all models emphasizes the importance of semantic similarity in financial summarization. Further experiments could investigate the alignment of BERTScore trends with human evaluation to validate its reliability as a metric.

7 Conclusion

Transformer-based models, such as T5 and Pegasus, excel in summarizing financial text, delivering a strong balance of accuracy and fluency. These models effectively extract key insights, making them highly valuable for financial analysis.The success of T5 and Pegasus demonstrates their suitability for abstractive summarization tasks in finance. Specifically, these models' ability to handle domain-specific language nuances and condense complex narratives into coherent summaries presents significant implications for financial applications.

Future research could focus on integrating multimodal summarization, incorporating non-textual data such as images, tables, and charts to offer richer insights for financial domains. Additionally, knowledge distillation can be employed to train smaller, efficient models that mimic the performance of larger models, ensuring cost-effective deployment. Reinforcement learning (RL) approaches hold potential to optimize summarization based on goals like informativeness and readability, enabling models to prioritize critical financial metrics. Finally, experimentation with large language models (LLMs) such as GPT may yield more sophisticated financial text summarization, leveraging their capability to adapt to complex domain-specific tasks.

References

[Abdaljalil and Bouamor2021] Samir Abdaljalil and Houda Bouamor. 2021. An exploration of automatic text summarization of financial reports. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing (FinNLP 2021)*, pages 1–11.

[La Quatra and Cagliero2020] Moreno La Quatra and Luca Cagliero. 2020. End-to-end training for financial report summarization. In *Proceedings of the First Financial Narrative Processing Workshop (FNP 2020)*, pages 171–180.

[Liu et al.2023] Shuaiqi Liu, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. 2023. Long text and multi-table summarization: Dataset and method. *arXiv preprint arXiv:2302.03815*.

[Mihalcea2004] Rada Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 170–173, Barcelona, Spain, July. Association for Computational Linguistics.

[Zhang et al.2024] Haopeng Zhang, Philip S Yu, and Jiawei Zhang. 2024. A systematic survey of text summarization: From statistical methods to large language models. *arXiv preprint arXiv:2406.11289*.