CS 584: SUMMARIZING WARREN BUFFETT'S LETTERS TO INVESTORS AND ANALYZING TRENDS OVER TIME

Katherine Shagalov* **and Aishwarya Bhethanabotla***
*Stevens Institute of Technology
kshagalo@stevens.edu, abhethan@stevens.edu
Spring 2025

ABSTRACT

This project demonstrates that combining abstractive summarization and topic modeling offers an effective strategy for analyzing long-form financial documents. Applying this dual approach to 45 years of Warren Buffett's shareholder letters, we successfully generated concise yearly summaries using a pretrained BART model and uncovered evolving thematic patterns through Latent Dirichlet Allocation (LDA). The quality of the generated summaries was evaluated using ROUGE metrics, while the topic modeling revealed temporal shifts in language and corporate focus across decades.

1 Introduction

The problem addressed in this study is the efficient summarization and thematic analysis of Warren Buffett's annual shareholder letters, spanning the years 1977 to 2021. These letters are widely regarded as a cornerstone of investment literature, offering decades of insight into Buffett's long-term philosophy, market observations, and capital allocation strategies. Despite their value, the cumulative length and complexity of the letters—often exceeding thousands of words per year—render manual analysis both time-consuming and impractical.

To address this challenge, we propose a natural language processing (NLP) approach that combines two complementary techniques: (1) abstractive summarization using large language models to condense each letter into a coherent and informative summary, and (2) unsupervised topic modeling to identify and track dominant themes over time. This dual-method framework enables a structured examination of how Buffett's perspectives on economic outlook, risk management, and corporate governance have evolved, particularly during periods of market turbulence such as the 2008 financial crisis and the COVID-19 pandemic. The findings of this study hold practical relevance for investors, financial analysts, and NLP researchers interested in long-form temporal document analysis.

2 Related Work

Recent advances in natural language processing have significantly improved the ability to summarize and analyze long-form texts. Early models like GPT[7] focused on left-to-right language modeling, limiting their contextual understanding. ELMo[6] addressed this partially by combining left-only and right-only representations but lacked deep interactivity between directions. BERT[2] introduced masked language modeling to capture bidirectional context, leading to strong performance on classification and understanding tasks. However, because BERT's predictions are not made auto-regressively, it is less suited for generation tasks like summarization. To bridge this gap, Lewis et al.[5] introduced BART, a denoising autoencoder that combines a bidirectional encoder with an autoregressive decoder, effectively aligning pre-training with generation objectives.

Other models like UniLM[3] and MASS[9] also proposed unified or span-based masking strategies for text generation, while XLNet[11] employed a permutation-based objective to capture bidirectional dependencies in an autoregressive framework. T5[8] further unified NLP tasks under a text-to-text format and achieved state-of-the-art results across several benchmarks through large-scale pretraining.

In parallel, topic modeling has remained a dominant approach for uncovering latent structure in document collections. Latent Dirichlet Allocation (LDA)[1] is still widely used, despite known limitations such as sensitivity to text

preprocessing and interpretability challenges. In the financial domain, prior work has focused on earnings calls and regulatory filings[4, 10], but relatively few studies have explored the unique narrative and temporal depth of shareholder letters. This project builds on these foundations by applying both abstractive summarization and topic modeling to a curated dataset of Warren Buffett's shareholder letters, aiming to extract high-level insights and track how key themes have evolved across economic cycles.

3 Methodology

We apply two complementary techniques:

Abstractive Summarization: We use the pre-trained facebook/bart-large-cnn model to summarize each letter. Due to BART's 1024-token input limit, we chunk long letters into 800-token segments, summarize each, and concatenate results.

Topic Modeling (LDA): We tokenize and preprocess the full letter texts, then train a Gensim-based LDA model to extract interpretable topics across the corpus. We filter out stopwords, punctuation, and domain-specific high-frequency terms (e.g., "berkshire").

We further assign dominant topics to each year and visualize their distribution to understand how Buffett's focus evolved.

4 Experimental Setup

4.1 Data

Dataset: 45 annual shareholder letters (1977-2021), each as a plain .txt file. We combined them into a pandas DataFrame with columns: year and letter_text.

4.2 Software and Settings

Python 3.11, Hugging Face Transformers, Gensim, NLTK, Matplotlib, ROUGE, and WordCloud. Code was run in Google Colab with GPU acceleration (T4).

Summarization Setup:

- Model: facebook/bart-large-cnn
- Input: 800-token chunks
- Output: Concatenated summaries per letter
- Saved results as buffett_letter_summaries.csv

LDA Setup:

- Model: Gensim LdaModel with num_topics=4, passes=25, iterations=300, random_state=3
- Preprocessing: Tokenization, stopword removal, frequency filtering (no_below=5, no_above=0.9)
- Custom stopwords: "berkshire", "hathaway", "would", etc.

4.3 Evaluation Metrics

ROUGE-1, ROUGE-2, ROUGE-L between original text and generated summaries. Topic visualization by year and top terms per decade.

5 Results

Summarization: BART produced high-quality summaries, retaining the financial tone and key points. Summaries were saved per year and evaluated using ROUGE.

ROUGE Scores:

- ROUGE-1: 0.5586

- ROUGE-2: 0.5240
- ROUGE-L: 0.5363

Topic Modeling: The final LDA model with 4 topics produced interpretable groupings:

- Topic 1: Scale, leadership, acquisitions (”billion”, ”ceo”, ”clayton”)
- Topic 2: Valuation & investment metrics (”intrinsic”, ”ratio”, ”bonds”)
- Topic 3: Subsidiaries & macro themes (”bnsf”, ”midamerican”, ”contracts”)
- Topic 4: Insurance & capital allocation (”float”, ”reinsurance”, ”contributions”)

Decade-Wise Top Keywords

Table 1: Top Keywords from 1980–1989

Word	Frequency
business	327
earnings	216
million	175
value	166
businesses	137
insurance	117
company	113
capital	108
market	108
companies	107

Table 2: Top Keywords from 1990–1999

Word	Frequency
business	249
earnings	212
value	188
company	188
million	178
stock	126
last	112
one	99
businesses	92
shares	88

Table 3: Top Keywords from 2000–2009

Word	Frequency
company	156
business	155
million	154
earnings	145
billion	124
value	122
last	120
one	97
insurance	90
may	82

Table 4: Top Keywords from 2010–2021

Word	Frequency
billion	241
earnings	211
value	167
company	163
business	147
million	116
businesses	110
many	106
last	105
shares	104

Word Cloud

Figure 1: Word Cloud of Buffett's Shareholder Letters

The word cloud highlights recurring themes in Buffett's letters, with dominant terms like "Berkshire," "year," "will," "business," "million," and "earnings."

BART vs. LDA Comparison

- BART summaries were readable and specific per letter.
- LDA captured broader latent themes and allowed trend analysis.

Sentiment Analysis

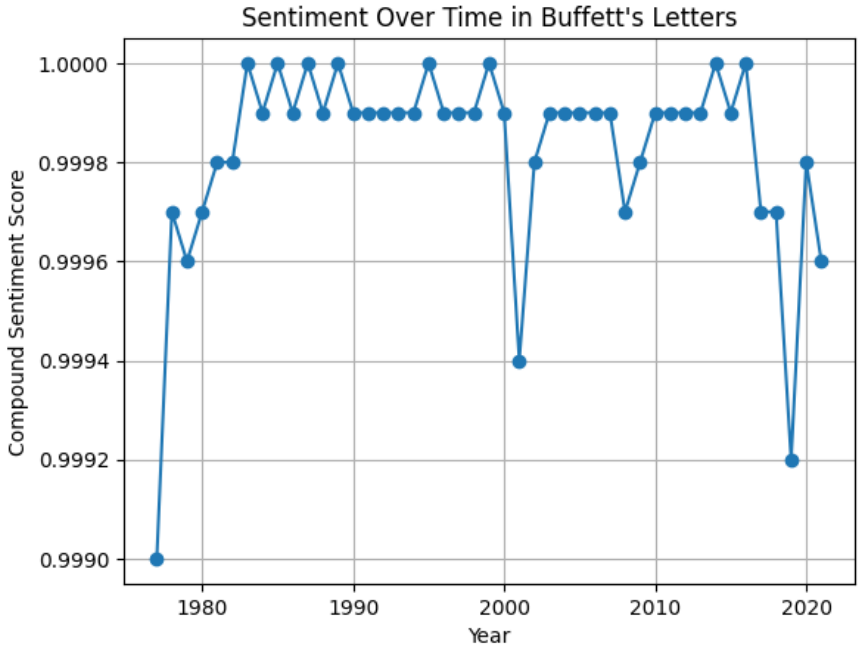

Figure 2: Sentiment Over Time in Buffett's Letters

Buffett's tone remained highly positive throughout, with minor sentiment dips during market downturns (2001, 2020).

6 Analysis

The dual approach of abstractive summarization and topic modeling proved highly effective in extracting and organizing insights from Warren Buffett's shareholder letters. BART provided readable, human-like summaries ideal for readers seeking quick understanding without diving into the full letter text. Its output successfully retained financial tone and highlighted key updates, particularly regarding acquisitions, capital allocation, and insurance performance. While BART slightly struggled with overly long or anecdotal letters (due to input length constraints), chunking and concatenation proved a sufficient workaround.

ROUGE evaluation scores were strong (ROUGE-1: 0.5586, ROUGE-2: 0.5240, ROUGE-L: 0.5363), especially considering the abstractiveness of the summaries and the length/complexity of the original texts. These scores suggest that the BART-generated summaries preserved significant content overlap with the source material, both at the unigram and bigram levels.

LDA topic modeling, in parallel, provided structure to Buffett's thematic focus across decades. Topics such as "insurance float," "reinsurance," and "billion-dollar acquisitions" align well with known historical milestones, such as Berkshire's expansion into utilities and transportation (MidAmerican, BNSF), or post-2008 conservative capital strategy. The topics were interpretable and distinguishable, with an even distribution of dominant themes across letters. This balanced spread indicates that Buffett's focus varied across different macroeconomic and corporate contexts.

Temporal analysis of keyword frequencies across decades reinforced LDA findings. In the 1980s, terms like "capital," "insurance," and "businesses" dominated. By the 2010s, keywords shifted to "billion," "value," and "shares," suggesting a broader corporate scale and capital growth focus. Word clouds visually confirmed these transitions.

Sentiment analysis further enriched our interpretation. Buffett's tone remained highly positive, even during downturns, underscoring his investor philosophy of long-term value and optimism. Minor dips in sentiment coincided with global financial disruptions (e.g., 2001, 2020), showing subtle tonal shifts.

Overall, the synergy of BART and LDA offered both micro-level readability and macro-level structure. Their alignment across lexical, thematic, and temporal dimensions validates the robustness of the approach. This method could easily be extended to other corporate communications to track leadership narrative, sentiment, and strategic evolution over time.

7 Conclusion and Future Work

This project demonstrates that combining abstractive summarization and topic modeling offers an effective strategy for analyzing long-form financial documents. Applying this dual approach to 45 years of Warren Buffett's shareholder letters, we successfully generated concise yearly summaries using a pretrained BART model and uncovered evolving thematic patterns through Latent Dirichlet Allocation (LDA). The quality of the generated summaries was evaluated using ROUGE metrics, while the topic modeling revealed temporal shifts in language and corporate focus across decades.

Notably, this work shows that even without fine-tuning, large language models like BART can produce interpretable and contextually rich summaries when paired with appropriate preprocessing techniques. Meanwhile, LDA proved valuable for surfacing high-level investment concepts, company-specific language, and the evolving strategic narrative of Berkshire Hathaway over time.

Looking ahead, several directions could enhance and expand this work. Incorporating extractive summarization methods such as TextRank would provide a useful baseline for comparison and help assess how well abstractive summaries retain core content. Fine-tuning BART on finance-specific corpora could further improve the contextual relevance and accuracy of the summaries. A more targeted, quantitative analysis of themes such as risk, leverage, or recession across different time periods could yield deeper insights into narrative shifts. Moreover, extending the framework to include sentiment or rhetorical analysis would allow for the detection of tone and communication strategy changes, particularly during times of economic uncertainty. Finally, applying this methodology to the letters of other high-profile CEOs—such as Jeff Bezos or Elon Musk—could enable meaningful comparisons across corporate leadership styles and strategic messaging.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[3] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*, 2019.

[4] Mahmoud El-Haj, Paul Rayson, Mark Walker, and William Young. Financial narrative summarization: Comparing nlp techniques for annual report summaries. In *Proceedings of the 2nd Financial Narrative Processing Workshop (FNP 2019), RANLP*, 2019.

[5] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

[6] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 2227–2237, 2018.

[7] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI Blog*, 2018. https://openai.com/research/language-unsupervised.

[8] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[9] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mass: Masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5926–5936, 2019.

[10] Zhi Xie, Xueming Liu, Di Wu, and Hui Xiong. Deep learning for financial text analysis: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[11] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32, 2019.