

Text mining financial statements: challenges and opportunities

Georgi Emilov Hristov¹

Abstract

This report examines the use of text mining approaches to analyze statutory financial statements, addressing key challenges like specific vocabulary and stop words. Through a case study on General Electric's consolidated statutory financial statement, it demonstrates the complexities that need to be managed in order to extract useful information from unstructured text (financial disclosures). The results demonstrate that using off-the-shelf, well known Python library (NLTK) is not sufficient when text mining statutory financial statements.

Ключови думи: text mining, financial statements, information quality

JEL: C38, M41

Introduction

In the rapidly evolving landscape of financial analysis, unstructured data, particularly in the form of textual information within statutory financial statements, presents both challenges and opportunities. Unlike structured data, which can be easily categorized and analyzed, the narrative portions of financial statements—such as management discussions, footnotes, and disclosures—often contain rich insights that are difficult to quantify. Text mining emerges as a powerful tool for extracting meaningful information from this unstructured text, enabling organizations to uncover trends, identify risks, and enhance decision-making processes. This report delves into the unique challenges posed by unstructured data in financial statements and explores the opportunities it presents for audit purposes, tax authorities, and analysts. By examining General Electric's consolidated statutory financial statement as a case study, the author highlights practical applications of text mining that can drive improved reporting accuracy and strategic insights in the financial domain.

This report provides an analysis which aims to determine whether existing functionalities provided by the NLTK Python library are sufficient for useful information extraction from financial statements. 2018 was a year of financial distress for General Electric². Its 2018 financial statements are analyzed to determine whether the results of this analysis provide the user with insight into the financial distress.

Discussion of Relevant Literature

The intersection of text mining and financial statement analysis has garnered attention in recent years, driven by the need for more sophisticated analytical tools to derive insights from large volumes of unstructured data. The literature reflects a growing recognition of the opportunities that text mining presents for financial analysts, auditors, and regulatory bodies, while also highlighting the inherent challenges associated with its application.

¹ PhD candidate, Department of Information Technologies and Communications, University of National and World Economics, Bulgaria, ORCID: [0000-0002-9641-5133](https://orcid.org/0000-0002-9641-5133), e-mail: georgi.hristov@unwe.bg

²<https://www.piranirisk.com/blog/general-electrics-financial-collapse-risk-management-case-study> (accessed: 01.11.2024)

Text Mining Techniques in Financial Analysis)

Numerous studies have explored various text mining techniques, such as natural language processing (NLP), sentiment analysis, and topic modeling, applied to financial texts. For instance, Loughran and McDonald (2016) emphasized the importance of adapting NLP techniques specifically for financial contexts, where language nuances can significantly alter interpretations. Their work suggests that traditional NLP methods may misinterpret financial jargon, potentially leading to inaccurate insights. Similarly, studies have demonstrated how sentiment analysis of management discussion and analysis (MD&A) sections can predict stock performance, reinforcing the value of integrating qualitative data into financial analysis (Li, 2010; Huang et al., 2020).

Challenges in Data Mining Financial Statements

Despite the promise of text mining, significant challenges persist. Data quality issues, including inconsistencies in formatting and terminology, complicate the extraction of meaningful insights. The impact of varying accounting standards, as discussed by Brown et al. (2014), highlights how differences in financial reporting can hinder comparative analyses across firms and industries. Moreover, the complexities of unstructured data, particularly in footnotes and narrative sections, require advanced analytical tools and methods to unlock valuable information (Beattie et al., 2004).

Practical Applications and Case Studies

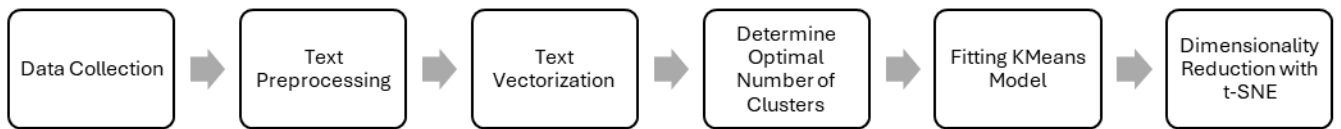
Practical applications of text mining in finance are increasingly documented. Case studies, like those conducted by Chen et al. (2019), illustrate how text mining can enhance traditional financial analysis by providing deeper insights into corporate governance, risk assessment, and investment decision-making. By analyzing financial statements alongside external data sources, researchers have shown that text mining can lead to more informed predictions of corporate performance (Zhang et al., 2021). The application of these techniques to General Electric's consolidated statutory financial statements serves as a pertinent example, demonstrating the potential for text mining to uncover hidden trends and facilitate more comprehensive financial evaluations.

Research Gap

The research gap identified in the existing literature lies in the limited understanding of how traditional text mining techniques, particularly without customized stop word filtering, may fail to capture nuanced insights in financial statements. While existing literature emphasizes the importance of integrating qualitative data into financial analysis, there is a lack of studies focusing specifically on the effectiveness of tailored text mining approaches in enhancing the interpretability of financial terms. This gap indicates a need for further exploration into developing customized methodologies that account for the unique language of finance and accounting, thereby improving the extraction of meaningful information from unstructured data in financial reporting.

Explanation of Methodology

To determine the challenges in text mining financial statements, General Electric's consolidated statutory financial statements for the 2018 financial year were sourced, preprocessed, and analyzed, following the pipeline, presented in Figure 1.



Source: Created by the author

Figure 1: Methodological Pipeline

Data Collection

The financial statements were sourced from General Electric’s official investor relations website, where they are published as part of regulatory requirements. The data typically comes in PDF format, which includes the annual report and quarterly filings. The choice of financial period (2018) is determined by the objective of the report.

Text Preprocessing

The preprocessing of the text extracted from financial statements involves several key tasks: it begins with extracting text from PDF pages, followed by segmenting this text into paragraphs. This process includes removing URLs, numbers, special symbols, and names, as well as applying lemmatization and eliminating stop words—both general and custom—relevant to financial terminology. This comprehensive cleaning ensures that the resulting text is uniform and focused, making it suitable for subsequent analysis.

Text Vectorization

The cleaned paragraphs are transformed into a numerical format using the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization method. This converts the text into a sparse matrix where each row represents a document and each column represents a term (word), with values indicating the importance of each term in relation to the document.

Optimal Number of Clusters

To determine the optimal number of clusters for the KMeans algorithm, both the Elbow Method and Silhouette Score are employed. The Elbow Method evaluates the inertia (the sum of squared distances of samples to their closest cluster center) for a range of cluster numbers (from 2 to 100). A plot of inertia against the number of clusters helps identify a "knee" point where adding more clusters yields diminishing returns. The Silhouette Score measures how similar an object is to its own cluster compared to other clusters, with scores closer to 1 indicating better-defined clusters.

The optimal number of clusters is found in two cases: one, using the generic English stop words list, provided by NLTK; and one using custom list of stop words, defined by the author (Appendix 1).

Fitting KMeans Model

Once the optimal number of clusters is determined, the KMeans clustering algorithm is fitted to the TF-IDF representation of the text data. The model assigns each document to one of the specified clusters.

Dimensionality Reduction with t-SNE

To visualize the clusters in a lower-dimensional space, t-distributed Stochastic Neighbor Embedding (t-SNE) is used. This technique reduces the high-dimensional TF-IDF data to two dimensions, allowing for easier visualization of cluster separations. The results are plotted to provide a visual representation of how documents are grouped.

Cluster Analysis

After clustering, the mean TF-IDF scores for each term in the clusters are calculated to identify the most significant words associated with each cluster. This analysis allows for interpretation of the clusters based on the prominent terms, providing insights into the themes or topics represented in the text data.

Data and Sources

The financial statements were sourced from General Electric's official investor relations¹ website, where they are published as part of regulatory requirements. The data typically comes in PDF format, which includes the annual report and quarterly filings. This ensures the analysis is based on the most recent official documentation available.

Objective of the Analysis and Assumptions

The objective of the analysis is to determine whether off-the-shelf, well-known NLTK functionality (a list of English stop words) is sufficient for the purposes of text mining statutory financial statements. To determine "sufficiency" several assumptions are made.

Information quality

A successful text mining task should provide the user with high quality information. Information is of quality if it is relevant – it has the potential to affect the user's decision making. Additionally, information should be understandable and timely. Since timeliness is highly dependent on the user, and the user's task, it is not assessed in this report.

Technicalities of relevance and understandability

For the purposes of the analysis the following additional assumptions are made:

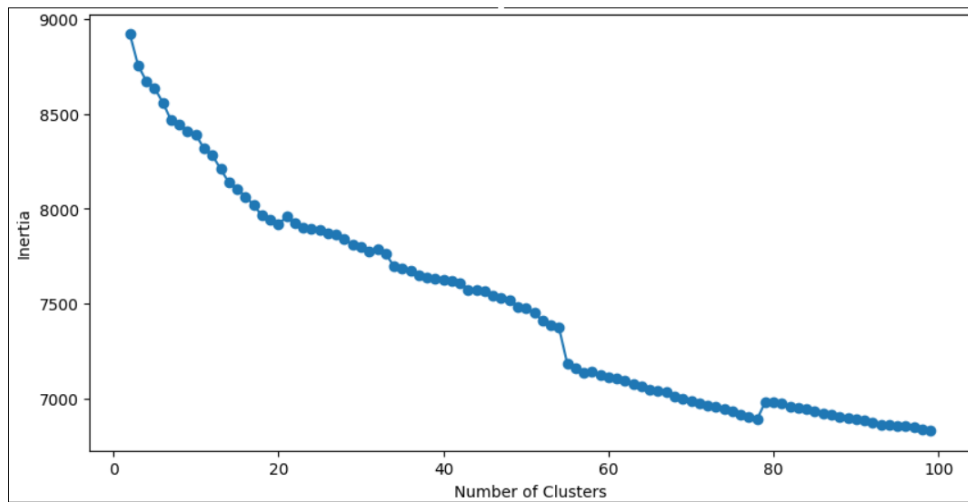
- The information from the analysis is inherently relevant, since the user is interested in all aspects of the results provided.
- The information from the analysis is understandable, if:
 - it forms less than 10 clusters, following the KMeans approach.
 - its content analysis (e.g. its TF-IDF scores provide the user with relevant words to consider).

Results

The Python code supporting this analysis is readily available from the author upon request.

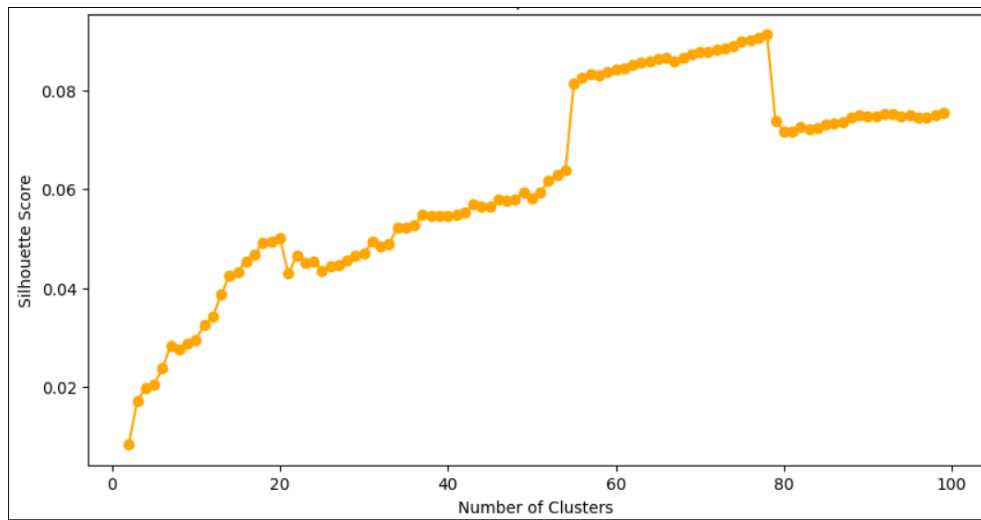
It was identified that the standard stop words provided by the NLTK library were insufficient for this analysis. After initial clustering with KMeans, the algorithm formed an unexpectedly high number of clusters (around 76 - determined by both Elbow method and Silhouette score), indicating a lack of homogeneity within the text data (Figure 2, Figure 4, and Figure 4). This result suggested that additional words, specifically those frequently appearing in financial statements but offering minimal analytical value (e.g., terms like "statement," "report," "financial," etc.), were influencing the clustering process and leading to unnecessary fragmentation.

¹ <https://www.ge.com/investor-relations/annual-report> (accessed: 01.11.2024)



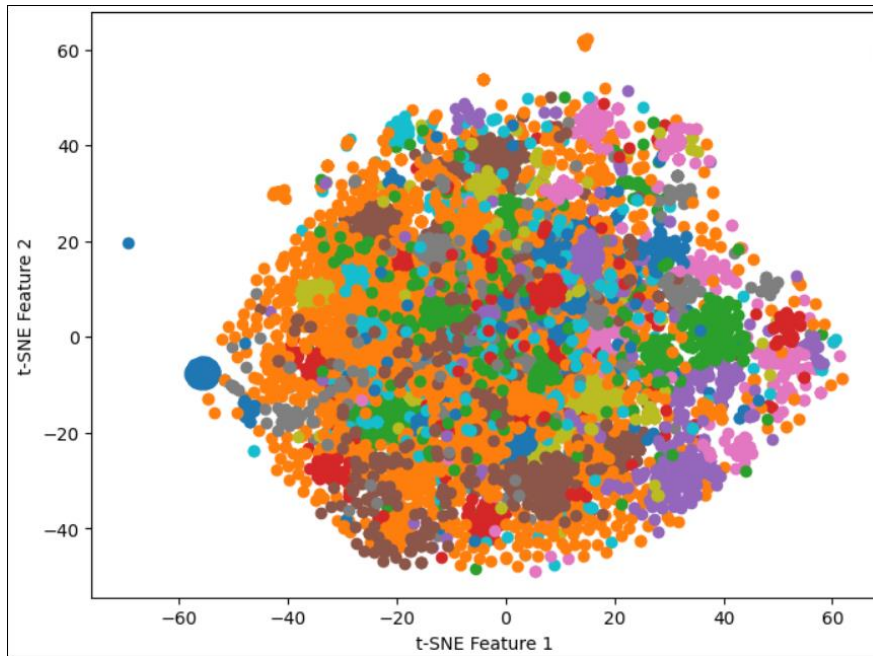
Source: Created by the author

Figure 2: Elbow Method for Optimal Number of Clusters (without custom stop words)



Source: Created by the author

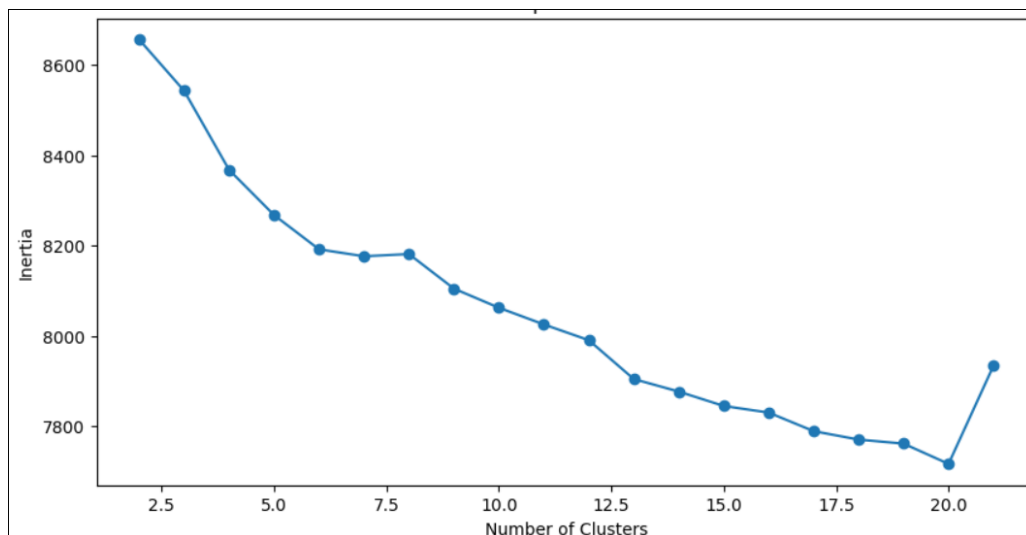
Figure 3: Silhouette Score for Optimal Number of Clusters (without custom stop words)



Source: Created by the author

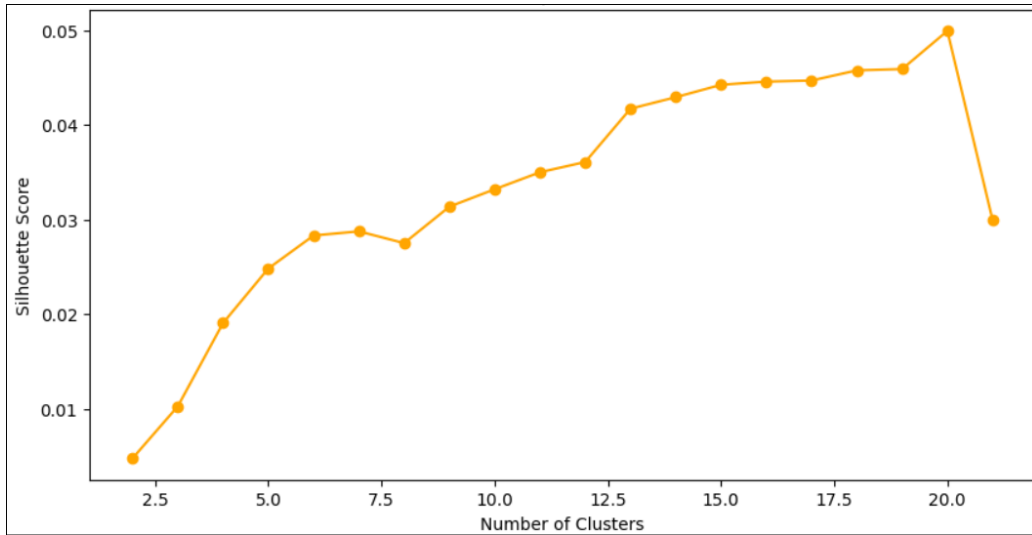
Figure 4: 76 KMeans Clusters (without custom stop words)

To address this, a custom list of stop words was created (Appendix 1) to capture these redundant terms. This adjustment significantly improved the clustering results by reducing the number of clusters (from 76 to 19), creating more coherent groupings that better represented distinct themes within the financial statements (Figure 5, Figure 6, Figure 7). The refined clusters now highlight the primary topics within the document, such as performance metrics, financial position, and operational details, allowing for a clearer and more actionable analysis of General Electric's 2018 financial performance.



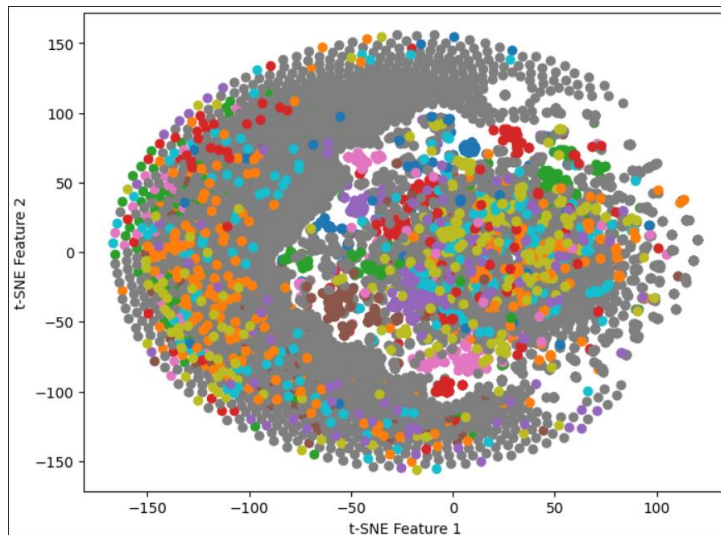
Source: Created by the author

Figure 5: Elbow Method for Optimal Number of Clusters (with custom stop words)



Source: Created by the author

Figure 6: Silhouette Score for Optimal Number of Clusters (with custom stop words)



Source: Created by the author

Figure 7: 19 KMeans Clusters (with custom stop words)

Additionally, analysis of the TF-IDF scores revealed that the most significant words were too common or generic, offering limited informational value. These terms lacked specificity, further reinforcing the need to

refine the stop word list to filter out high-frequency, low-value words commonly found in financial statements.

Significant words without a custom list of stop words: *cost, expense, benefit, service, operating, statement, financial, consolidated, note, information, position, section, derivative, audited, cash, restricted, equivalent, flow, hedge*

Significant words with custom list of stop words: *goodwill, borrowing, impairment, loan, adverse, decrease, discontinued, earnings, provision, obligation, reserve, generated, activity, pricing, impact, treasury, private, disposition*

The comparison of significant words with and without a custom stop word list shows that tailored filtering improves analytical precision. Without custom stop words, the analysis yields general financial terms like "cost" and "expense" that are too generic to provide specific insights. However, with a customized list, terms such as "impairment," "discontinued," and "provision" become more prominent, highlighting more specific financial activities and conditions (specifically financial distress). This refined focus allows the analysis to reveal deeper insights into areas like financial health and risk management, making it more useful for stakeholders such as auditors, analysts, and tax authorities.

Limitations of the Analysis

The limitations of this report include the narrow focus on General Electric's financial statements, which may not capture all relevant financial data. The custom stop word list used is subjective (created by the author) and may lead to missed insights or retained irrelevant terms. The analysis depends on the accuracy of natural language processing tools, which may struggle with financial jargon. Additionally, the findings may not be generalizable to other companies or industries and are limited to a specific time frame (2018), potentially overlooking long-term trends. These factors highlight the need for further research to refine methodologies for broader applicability.

Conclusion

This report demonstrates the potential of text mining to extract meaningful insights from unstructured financial data, such as General Electric's 2018 financial statements. The analysis highlighted several challenges, particularly in the preprocessing stage, where the presence of common financial terms diluted the effectiveness of traditional stop word lists. By incorporating a customized list of stop words tailored to the language of financial reporting, it was possible to enhance cluster coherence and reduce redundancy.

The refined clusters offered clearer themes, revealing key aspects of the company's financial and operational narrative (financial distress). However, the initial issues with clustering and TF-IDF analysis underscore the complexities inherent in text mining financial documents. Future studies could benefit from further refinement of preprocessing methods, including techniques that adapt dynamically to context-specific jargon. Overall, the findings affirm the value of text mining as a tool for analysts, auditors, and regulatory bodies to better understand and interpret extensive, complex financial documents. The code used in this analysis is available from the author upon request, supporting reproducibility and encouraging further exploration in this field.

References

Beattie, V., McInnes, B., & Fearnley, S. (2004). A methodology for analyzing and evaluating narratives in annual reports: A comprehensive descriptive profile and metrics for disclosure quality attributes. *Accounting Forum*, 28(3), 205–236. <https://doi.org/10.1016/j.accfor.2004.07.001>

Brown, P., Preiato, J., & Tarca, A. (2014). Measuring country differences in enforcement of accounting standards: An audit and enforcement proxy. *Journal of Business Finance & Accounting*, 41(1–2), 1–52. <https://doi.org/10.1111/jbfa.12066>

Chen, H., De, P., Hu, Y. J., & Hwang, B. H. (2019). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, 27(5), 1367–1403. <https://doi.org/10.1093/rfs/hhu082>

Huang, A. H., Zang, A. Y., & Zheng, R. (2020). Evidence on the information content of text in analyst reports. *The Accounting Review*, 95(4), 271–304. <https://doi.org/10.2308/accr-52669>

Li, F. (2010). The information content of forward-looking statements in corporate filings: A naïve Bayesian machine learning approach. *Journal of Accounting Research*, 48(5), 1049–1102. <https://doi.org/10.1111/j.1475-679X.2010.00382.x>

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4), 1187–1230. <https://doi.org/10.1111/1475-679X.12123>

Zhang, J., Xie, Y., & Zhang, Y. (2021). Textual analysis and firm valuation: A machine learning approach using financial filings. *Journal of Financial Economics*, 142(1), 122–140. <https://doi.org/10.1016/j.jfineco.2020.11.001>

Appendix 1 – Custom List of Stop Words

account, accounts, accounting, accrual, accrued, asset, assets, balance, capital, cash, consolidated, cost, debt, expenses, financial, financials, forecast, income, liability, liabilities, management, net, operating, report, reports, revenue, statement, statements, tax, total, years, year, yearly, due, company, business, transactions, results, gains, losses, profit, profits, loss, financially, liquidation, valuation, audit, auditor, ratios, equity, equities, interest, return, returns, current, previous, future, period, periodic, quarter, annual, assess, analysis, assessment, entity, policy, policies, reporting, prepare, prepared, require, requirements, businesses, services, service, trade, million, billion, electric, general, incorporated, reference, corporation, form, agreement, exhibit, dated, amended, executive, effective, stock, ge, restated, segment, industrial, mda, operation, margin, sub, gas, dollar, ended, healthcare, lighting, aviation, transportation, power, oil, energy, renewable, steam, generation, solution, equipment