Automated Qualitative Summarization of Financial Statements and Corporate Disclosures

Joshua Bae, Aneel Damaraju, Andrew Pham, Kunal Rai, Angel Soto

1. Introduction

Financial data and company information is often hidden behind layers of fancy terminology, numeric ratios, and the guise that financial advisors are needed in order to make good investment decisions. The SEC (U.S. Securities and Exchange Commission) requires companies to disclose this information in their quarterly reports in a document formally titled the 10-Q, but these are often over 50 pages and are not written for the average person. Through automatic qualitative summarization of financial statements and corporate disclosures, we would like to provide easy-to-understand summaries through tagging and extracting important and useful qualitative sentences in the financial reports.

We extract the most useful information from these reports by building a model where important sentences for determining company performance are extracted from the quarterly report. While each report is structured in the same general way, the content in each of the subsections of the report can vary depending on the activities of the company in the previous quarter. For example, executives tend to keep the market risks of their company as brief as possible, but sometimes, there is information of value in the risk section. In addition, we can compare companies within a sector of the economy, basing valuable sentences from a combination of what is often talked about within a certain market.

If we can identify the key sentences in a quarterly report and extract them, it would then be possible to algorithmically generate a summary of quarterly reports that is standardized and more approachable.

Project Description

This project will focus on the identification of key information in quarterly reports. While executive/management summaries of financial reports do detail some of the performance metrics and outlook of the company, much of the important information, related to fundamentals, earnings, and projections, are hidden behind acronyms, ratios, and terminology that makes it increasingly difficult for a layperson to extract key information. This project will utilize quarterly reports from 10-Q statements to identify sentences that explain the current state of the company and indications of the future actions by the company. The 10-Q consists of large amounts of accounting statements, legalese, and required disclosures, but the data can be mined for a more concise qualitative summary of the company.

Domain Knowledge

Barriers to comprehension of the 10-Q report come from its length, long legal statements, financial syntax, and uneven dispersion of important information throughout the document. In order to understand all of the items in the 10-Q report, basic knowledge of the following is needed: financial statements and ratios, common short-term business decisions, industry knowledge of the specific company, importance of products to business performance, key performance metrics of businesses, and risk factors. Since the scope of this project is limited to qualitative information from the 10-Q, financial accounting will be removed from the required domain knowledge. For information that is commonly gleaned from the 10-Q by investors, which includes product performance, industry performance, and short-term business decisions, the Management's Discussion and Analysis (typically Part 1 Item 2 of the report) is important. This section has some boilerplate text but includes a section where the CEO or upper management explains the plan for the next quarter or longer term for the business. The Management Discussion also typically includes the target customer profile, how the business is reducing costs, and how the company plans to continue its growth (growth strategy). This discussion of the plan for the company and the potential growth is important for any investor to understand. In conjunction with this part, Part II of each 10-Q contains a section on risk assessment. This section entails any of the things that could possibly go wrong for the business and how it could negatively impact the business. Although this section is terse, a few of the sentences provide value to an investor as it provides potential risks and downside. It is important that the extractive summary include key sentences from these sections as it will give a comprehensive overview of important information to the layperson to understand.

Prior Work

In the field of text summarization, there is a significant amount of research in extractive and abstractive techniques. In our survey of prior work, we focused on extractive and finance related work. Extractive techniques have tended to be based on five main methods: Term frequency-inverse document frequency (TFIDF), clustering, graph based, latent semantic analysis (LSA), and more recently, neural networks [1]. Clustering and graph-based approaches are mainly unsupervised methods that use vectorized representations of the documents to determine different sentences and sentence importance. LSA focuses on semantic relationship, which does

not require that words match but takes the conceptual relationships between documents using singular value decomposition (SVD) on document-word matrices. Neural networks are typically applied through a supervised method where human proof-read and marked documents can be used to train a network to identify which sentences are deemed as important for specific domain but falls short in the case that no human cleaned text is available to train on. As supervised learning requires a training set, semi-supervised learning is a more commonly used middle ground since it achieves comparable results to supervised learning without the same training time required [2]. In many of these methods, sentence representations in word or sentence embeddings are used to make the information interpretable by learning models. In the forefront of embeddings, Google's BERT has been a top performer using a transformer model that is bi-directional and has a language model that looks forward and backwards. BERT models are pre-trained and allow for use directly in a pipeline for embeddings. Through this, FinBERT a BERT language model developed for finance related NLP, was created. FinBERT improved on state-of-the-art results for finance related tasks and outperforms other learning methods for the finance domain [3].

Note:
For the purpose of this report and in the completion of this project, the models described above were attempted to be used but did not meet the scoring standards of the final model chosen (detailed in the *Modeling* section). We reached out to the team that created FinBERT to see if we could use their model for our project, but they were unable to share the model with us.

Objectives

Our two major project objectives are as follows:

a. Obtain quarterly reports from well-known tech corporations relevant to the project and clean and structure the data.
b. Identify the most important information for summary generation and utilize NLP to extract the important sentences.

2. Data Description

The SEC requires that all companies file financial reports in a standardized text format, such as the 10-Q format for quarterly reports, and the 10-K format for annual reports. The 10-Q follows a specific itemized format: Financial Statements, Management's Discussion, Risk Assessment, and Controls and Procedures. However, companies can add more sections on their own discretion. As an example, Figure 1 denotes the table of contents of Apple Inc.'s second-quarter 10-Q from 2018.

Apple Inc.

Form 10-Q

For the Fiscal Quarter Ended June 30, 2018

TABLE OF CONTENTS

Figure 1: The standardized sections of an Apple Inc. 10-Q report, Q2 2018.

Revenue

Revenue decreased $12 million, or 0.6%, during the three-month period, due to (1) the reduction in revenue from the sale of the Capco consulting business and the risk and compliance consulting business during the third quarter of 2017 and (2) the sale of the Certegy Check Services business unit during the third quarter of 2018. These decreases were partially offset by (1) growth in retail payments; (2) increased volumes in banking and wealth solutions (excluding the effects of the sale of the risk and compliance consulting business); (3) growth in GFS banking and payments solutions in North America; and (4) payments growth in Latin America. Additionally, the three months ended September 30, 2018 was also impacted from a $30 million unfavorable foreign currency impact primarily resulting from a stronger U.S. Dollar in 2018 versus the Brazilian Real.

Figure 2: A section of Fidelity National Information Services, Inc.'s 10-Q report. Note the highlighted phrase is of value since it notes the reason for a change in revenue of the company.

In Figure 2, Fidelity National Information Services' Q3 2018 report, the phrase "due to (1) the reduction in revenue from the sale of the Capco consulting business and the risk and compliance consulting business during the third quarter of 2017 and (2) the sale of the Certegy Check Services business unit during the third quarter of 2018", is something that should be flagged as it details industry trends and how they are managing and accounting for a decline in sales performance.

RESULTS OF OPERATIONS

Financial Performance Summary for the First Quarter of Fiscal 2018

- Total Digital Media ARR of approximately $5.72 billion as of March 2, 2018 increased by $336 million, or 6%, from $5.39 billion as of December 1, 2017. The change in our Digital Media ARR is primarily due to strong adoption of our Creative Cloud and Adobe Document Cloud subscription offerings.

Figure 3: Example of Management Discussion (qualitative overview) Part 1, Item 2 of ADOBE SYSTEMS INC.'s 2018 1st Quarter 10-Q report. The highlighted sentence is of value since it contains information on a new subscription implemented by the company that raised revenue.

Shown in Figure 3 is Adobe Systems' Management Discussion section from their first quarter 10-Q from 2018. In this discussion, the report highlights their "adoption of [their] Creative Cloud and Adobe Document Cloud subscription offerings". We believe that subscription or acquisition announcements like these could potentially be correlated with strong current performance and indicative of continued prosperity in the future.

Data Exploration

First, we manually examined fifteen randomly selected 10-Q reports and manually created extractive summaries for each of these documents. These fifteen manually generated summaries will later be split into two groups: one group for training any supervised models we may use and one group for validation.

In this process we established the types of phrases that were important, identified key characteristics included in each report, and found sections of the report that contain mostly relevant information to our interests. For types of phrases that we found important, we established that we would not want any company projections (regarding their own forecasting) but were interested in how the company planned their growth strategy moving forward. Additionally, sentences that provided information about the general products and services that the company provided were also deemed as important. This was done so that anyone who reads the sentences extracted could get a general overview of the company and their offerings.

To determine which items had important qualitative information, we have summarized the content that is presented in each section of any 10-Q.

Table 1: Description of Sections of a 10-Q

Item 1, *Financial Statements* (Part 1)	Covers financial statements and various legal proceedings. Much of the information here is quantitative financial data.
Item 2, *Management's Discussion and Analysis of Financial Conditions and Results of Operations* (MD&A) (Part 1)	Rich in valuable information as it pertains to the management's perspective on the company performance. The company analyzes their quarterly performance and provides information on their growth strategy and how they are handling losses. Future goals and approaches to new projects are also discussed in this section.
Item 3, *Quantitative and Qualitative Disclosures About Market Risk* (Part 1)	Informs readers of potential exposure to loss arising from changes in interest rates, foreign currency exchange rates, etc.
Part 2 of 10-Q (*Legal Proceedings* and *Risk Factors*)	Beyond *Legal Proceedings* and *Risk Factors*, much of Part 2 is primarily left blank as it does not pertain to many of the companies included in the data set.

Since the important qualitative information is spread across the entire 10-Q with different density in each section, we determined that we needed to split documents across these items.

Figure 4 depicts a sentence that we deemed as important. Sentences like these are important as they convey information about the future of the company and what the company is working on in a qualitative manner.

The Company announced iOS 13, macOS® Catalina, watchOS® 6 and tvOS® 13, updates to its existing operating systems, and introduced iPadOS™, a new operating system designed for iPad, all of which are expected to be available in the fall of 2019. The Company also introduced an updated Mac Pro® and an all-new Pro Display XDR, both of which are expected to be available in the fall of 2019.

Figure 4: A sentence of value, taken from an APPLE INC. 2019 3rd Quarter 10-Q. The sentence describes new Apple products and software updates which is important information about the outlook of a company.

We also examined word relevance to explore the viability of keyword extraction methods by creating TFIDF matrices for a subset of the documents.

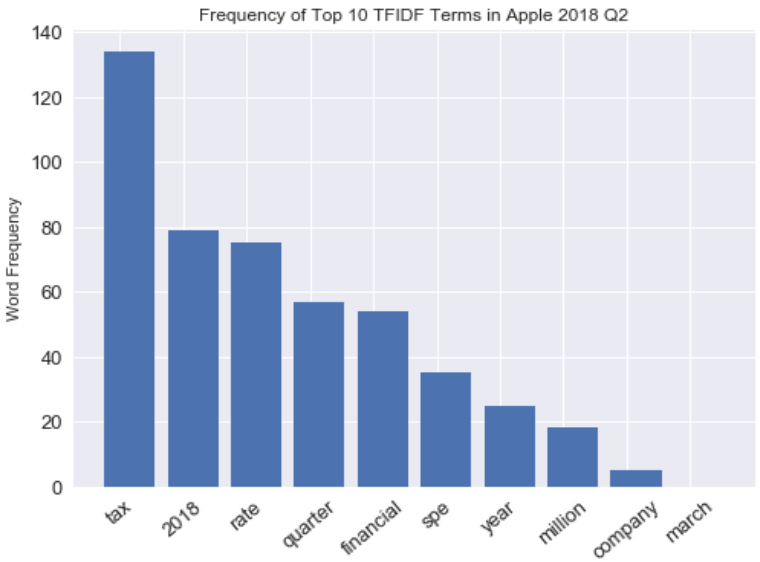
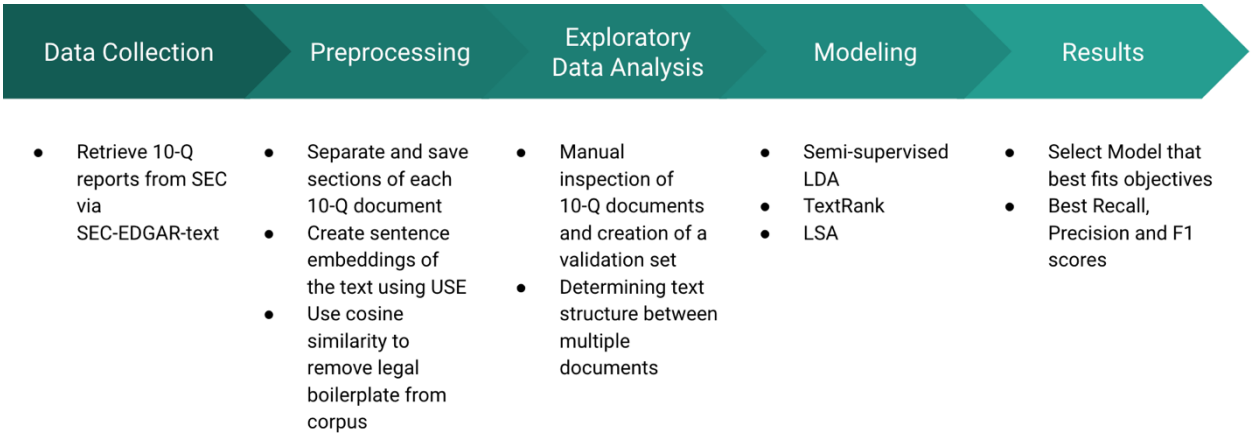


Figure 5: TFIDF terms of Apple 2018 Q2

Figure 5 depicts the TFIDF for Apple 2018 Q2. The TF-IDF terms are not particularly useful or indicative of relevant sentences. Thus, we were able to conclude that automated keyword extraction would not be a successful approach for this project, and that more complex methods outlined in the *Modeling* section would be needed.

3. Methods & Models



Data Wrangling

Financial statements can be obtained from the SEC via cURL or Python *urllib* requests through their EDGAR platform. We initially had written scripts to output the HTML of the quarterly report given a company and the quarter to automate data collection, however we found HTML from the SEC to be unstructured and difficult to parse. We were able to find a repository made by Bill McDonald (Professor of Finance at Notre Dame, alions7000) that took the json and eXtensible Business Reporting Language (XBRL) files from EDGAR and converted them to

plain text for parsing. For 62 companies in the teach industry, we acquired six 10-Q reports from December 2017 to December 2019 (in .txt format) via *SEC-EDGAR-text* [4].

We then took this easily formattable text data and split the text files by the section headers to allow for more detailed analysis and control over the data present in our corpus. The output of this data wrangling step was 5 text files corresponding to the 5 sections for each 10-Q.

Finally, as a method of reducing biases in creating the most accurate model, we decided to go with an 80/20 split in which 80% of our 10-Qs would be used to train our models and 20% would be set aside to later validate the performance of our models. To do this, we randomly selected 20% 10-Q (approximately 140 statements) and set them aside for testing. We took extra precaution in not using this information to influence any of our decisions for the pipeline (blinding).

Preprocessing

Once the data was processed into a readable and easy to analyze format, we worked to make the data easier to process quantitatively. To do this, we created fixed-length vectors (i.e. sentence embeddings) of size 750 for each sentence using the Universal Sentence Encoder (USE), which captures the semantic and syntactical structure of a sentence in this vector. Creating these fixed-length vectors allowed us to easily perform semantic similarity comparisons between sentences with a simple inner product between their corresponding embeddings.

Removing Legal Boilerplate

Companies are required by the SEC to include certain legal and tax-related cautionary statements, and these sentences are dispersed throughout the 10-Q. Below is an example of one of these sentences from the 2018 Q3 Apple 10-Q.

Beginning in 2018, the Company records any excess tax benefits or deficiencies from its equity awards as part of the provision for income taxes in its Condensed Consolidated Statements of Operations in the reporting periods in which equity vesting occurs.

However, we do not want these sentences to be included in the final summary. Removing this legal boilerplate before we even reach the modeling stage would be the most effective way to achieve this goal. Through the excerpts in the paper *Litigation Risk and Voluntary Disclosure: The Use of Meaningful Cautionary Language,* we obtained a set of legal and tax-related cautionary statements that often appear in financial statements [5].

As an example, we generated sentence embeddings of the boilerplate sentences and the sentences from a 2018 Q1 Apple 10-Q. From these embeddings, we constructed a cosine similarity matrix displayed in Figure 6.

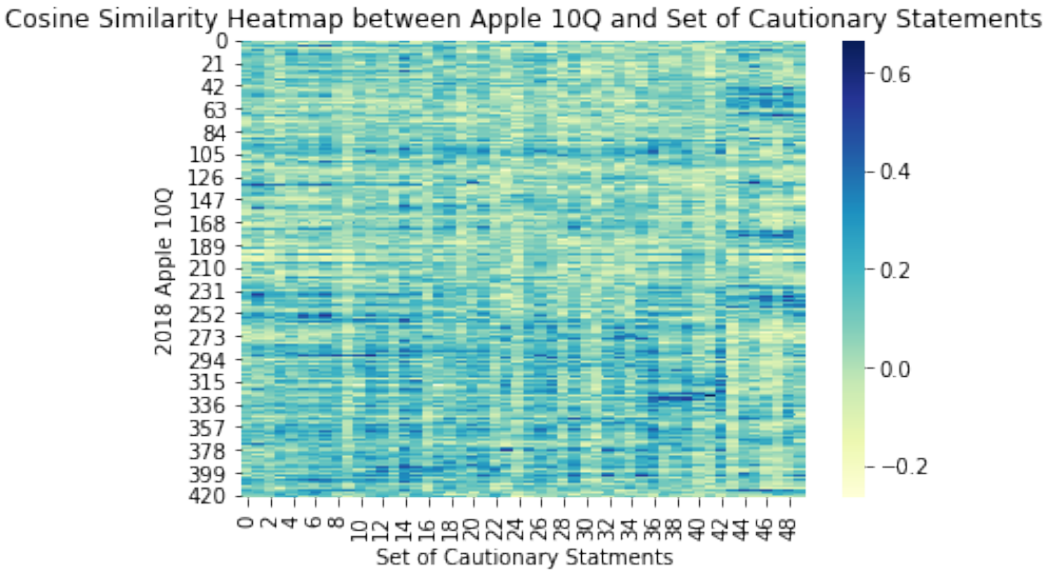


Figure 6: Cosine Similarity Matrix Between Apple 10-Q and Set of Cautionary Statements

From the dispersion of blue spikes throughout the heatmap, we observed that the legal sentences are dispersed quite broadly throughout the entirety of the 10-Q. Furthermore, we compared the distribution of legal sentence similarity scores between the legal bank and four different companies in Figure 7.

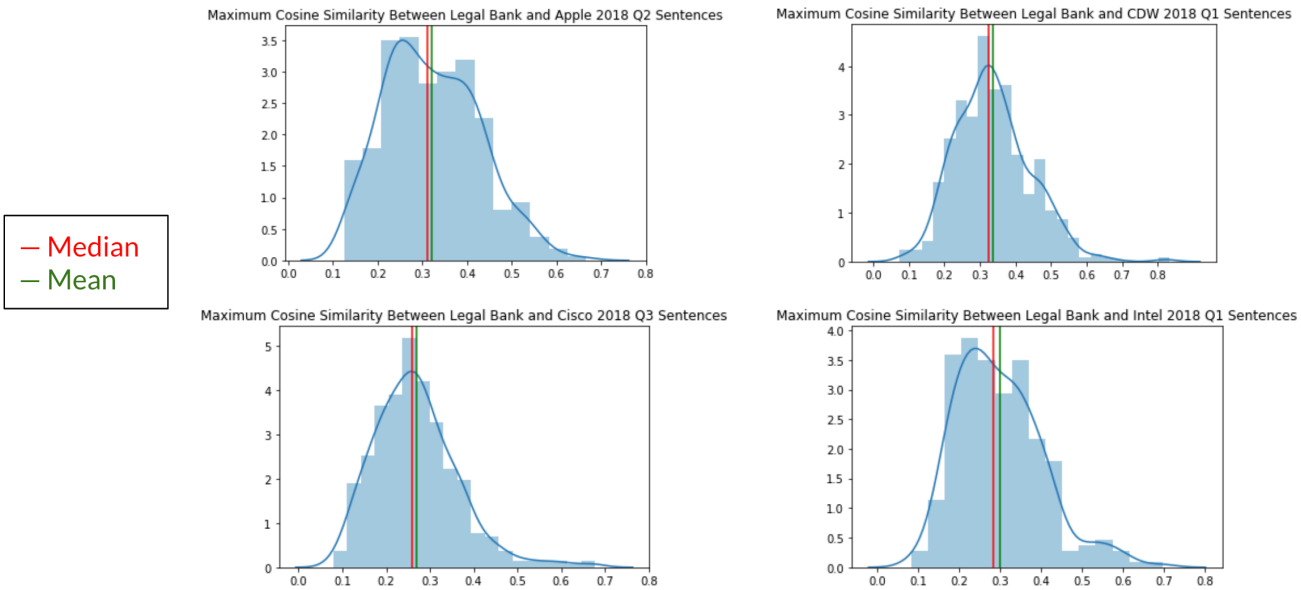


Figure 7: Distribution of Legal Sentence Similarities for Apple, CDW, Cisco, and Intel

The means and medians of the legal similarity scores are highly variable between companies. Since variation in how formally a company writes its documents affects the similarity score, we decided that choosing a global threshold was not the optimal solution. Instead, the removal

threshold for each document would be one standard deviation above the document's mean. This would allow the threshold to be more tailored to the specific document.

By selecting this "one standard deviation above" threshold, we removed remove legal boilerplate quite effectively through a purely statistical process. For example, by using a threshold of 0.4 for the above Apple 10-Q, we were able to extract sentences such as "*Due to economic and political conditions, tax rates in various jurisdictions may be subject to significant change*" while retaining sentences unrelated to tax or legalities. By passing the sentences from each document through this legal boilerplate detection algorithm, we effectively reduced the amount of noise (boilerplate sentences) in our data before the modeling stage.

Pre-Modeling

During our data exploration phase, we noticed that a disproportionate number of useful qualitative sentences came from the *Item 2. Management's Discussion and Analysis of Financial Condition and Results of Operations* section. When we manually generated the summaries for a 10-Q, we found that on average, we took 20% more sentences from this section relative to its proportionate length.

Therefore, for every model, we decided to run each of the five sections of the 10-Q through the model individually. Then, we imparted a bias on every model to take 20% more sentences from *Item 2: Management's Discussion* compared to how many we would have taken if we had chosen purely based on its proportion to the length of the section.[1] As a result, fewer sentences were pulled from sections deemed less important (to satisfy our sentence quota). This will allow the model to replicate selection tendencies we used in generating the manual summaries and increase the model scores.

Modeling

Common approaches to text summarization include semi-supervised and unsupervised models. Below, we will outline our two best performing models: Latent Dirichlet Allocation (LDA) modified for semi-supervised learning and TextRank [6]. After passing the sentence embeddings of the 10-Qs through our preprocessing stage to remove legal and tax-related boilerplate, we then input our remaining embeddings into both approaches to generate extractive summaries.

The Appendix outlines a complete list of all the models we tried, a brief description of each, and their performance results.

[1] We imparted the bias by creating a higher-order function inside *modeling/run_model.py* that takes a model implementation as a callable function. Then, for a given 10-Q, the higher order function will run each section through the model and take 20% more sentences from the model output on *Item 2. Management's Discussion.* The number of sentences to take for each section (with the bias included) is calculated in *modeling/compute_num_cluster.py*. This implementation approach allowed us to easily generate summaries for any model we created and consistently apply the bias to each model to obtain similar-length summaries.

TextRank

TextRank is an unsupervised, graph-based approach based on the PageRank algorithm, which constructs a graph with webpages as nodes and hyperlinks as directed edges and assigns weights to webpages based on some importance metric. TextRank replaces webpages with sentences as the nodes of the graph and retains the PageRank logic to assign importance scores to the sentences [6]. For our project, sentences with the highest scores of importance make up our summary. A sample graph for sentence extraction is shown in Figure 8.

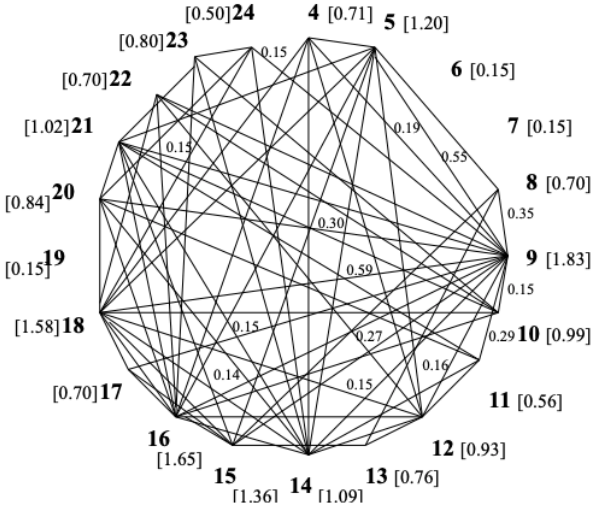


Figure 8: Sample Graph for TextRank, where each node represented by a bolded number represents a sentence. The numbers on the edges of this graph represent the similarity between the two nodes. The sentence similarity score in brackets ranks the sentences.

One possible advantage to this graph-based approach is that it does not require a deep understanding of linguistic or domain-specific knowledge. However, given that financial documents have a significant amount of complex vocabulary, TextRank may be a convenient but suboptimal summarization technique.

Semi-Supervised Latent Dirichlet Allocation (LDA)

LDA was used as an unsupervised topic model that classifies text in a document into particular topics, with the number of topics being a model hyperparameter [7]. These topics consist of linear combinations of different keywords contained in a sentence, with the weight of each keyword corresponding to its importance to the respective topics. A sample breakdown of a topic can be seen in Figure 9. Once this topic model is created, each sentence is passed through the trained model, resulting in a "topic vector" with length equal to the number of topics. The topic vector is a normalized vector, where the value of the vector at index *i* indicating how much of the sentence falls into topic *i*. Essentially, LDA is another way to vectorize sentences, but is finetuned to the dataset that we are using, since we can train the model on all of the documents in our training set.

We also use LDA as a crude modeling technique, picking sentences that fall mostly into a single topic as those that are maximally relevant. This model is used mostly as a comparison metric, as can be seen in the results section.

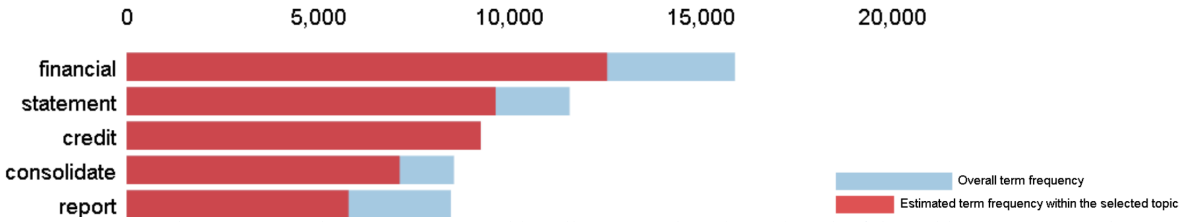


Figure 9: A sample topic created using Latent Dirichlet Allocation. The relative length of the red bars represents their importance to the topic, and the blue bars represent their frequency in the document.

Once the model has been trained and the sentences have been vectorized, we propose using these vectors to create a novel semi-supervised algorithm that ranks sentences based on cosine similarity. First, we create a validation set of sentences that would be candidates for sentences that would come from a reasonable summary and vectorize them with LDA. Then we take the average of these vectors, creating a target topic vector that we then compare to every sentence in the training set. Sentences with the largest cosine similarity are considered good summary sentences and included in the summary.

While this method has room for improvement, this novel semi-supervised approach is useful since it incorporates an LDA model finetuned on the 10-Q reports. It is also trained on sentences that we consider important in a summary, which allows us to include domain specific information in the model training process. For this project, we choose the number of topics to be 15, as the keywords created when we used 15 topics seemed to capture the various main sections of the 10-Q.

4. Results

While nine of the fifteen manually generated summaries were used for training the Semi-Supervised LDA model, the remaining six were used as a validation set for our models. To assess our models, we assessed the performance of the model on the validation set, using the manually generated summaries for the validation set as the "truth". We used the precision, recall, and F1 metrics to systematically assess our models. In the following metrics, the phrase *correctly retrieved sentences* refer to the set of the sentences that the automated summary pulls out that matches a sentence in the corresponding manually generated summary.

Precision

Precision measures the proportion of correctly retrieved sentences out of total sentences in the generated summary.

Validation Filename	Random	TextRank	Unsupervised LDA	Semisupervised LDA
Apple 2018 Q2	0	0.1	0.1	0.1
Vishay IT 2018 Q1	0	0.1	0.2	0.5
IBM 2018 Q1	0	0.0625	0	0.125
Applied Materials 2018 Q2	0	0.2	0.1333	0

Cisco 2018 Q3	0	0.0344	0	0.0689
Mantech 2018 Q1	0.0833	0.333	0.1666	0.0833
Mean	0.0138	0.138	0.1	**0.1462**
Median	0	0.1	0.116	0.091
Standard Deviation	0.034	0.1107	0.0843	0.1783

Recall

Recall measures the proportion of correctly retrieved sentences out of total number of sentences in the manually generated summary.

Validation Filename	Random	TextRank	Unsupervised LDA	Semisupervised LDA
Apple 2018 Q2	0	0.066	0.0625	0.0769
Vishay IT 2018 Q1	0	0.0625	0.125	0.3125
IBM 2018 Q1	0	0.055	0	0.111
Applied Materials 2018 Q2	0	0.1875	0.125	0
Cisco 2018 Q3	0	0.0625	0	0.1428
Mantech 2018 Q1	0.076	0.25	0.125	0.0625
Mean	0.0128	0.114	0.0729	**0.117**
Median	0	0.0645	0.093	0.094
Standard Deviation	0.031	0.083	0.061	0.107

F1

F1 is a metric that combines the recall and precision scores via the formula $F_1 = \frac{2*precision*recall}{precision+recall}$ to present a single metric for a summary's performance.

Validation Filename	Random	TextRank	Unsupervised LDA	Semisupervised LDA
Apple 2018 Q2	0	0.08	0.076	0.087
Vishay IT 2018 Q1	0	0.076	0.154	0.385
IBM 2018 Q1	0	0.0588	0	0.118
Applied Materials 2018 Q2	0	0.193	0.129	0
Cisco 2018 Q3	0	0.044	0	0.093
Mantech 2018 Q1	0.08	0.285	0.143	0.074
Mean	0.0133	0.123	0.083	**0.126**
Median	0	0.078	0.103	0.0899
Standard Deviation	0.032	0.095	0.07	0.133

In every metric, the semi-supervised LDA model had the highest mean score across the validation documents. In a document-by-document comparison of the F1 scores for semi-supervised LDA and TextRank (the second best-performing model), we can see that semi-supervised LDA typically performs better. However, for Vishay IT, semi-supervised LDA performs significantly better while it scores a 0 on Applied Materials. A small number of score fluctuations like these increases semi-supervised LDA's performance variability.

5. Conclusion

With our model being able to extract our human selected sentences with a mean precision score of 0.1462, our semi-supervised approach surpassed unsupervised approaches such as Latent Semantic Analysis and TextRank. By including the unsupervised LDA as a point of comparison, we can conclude that the domain knowledge that we incorporated into our novel semi-supervised LDA model significantly improved model performance. This was especially helpful in the context of financial documents, where obscure terminology is not significantly correlated with sentence relevance.

However, with a limited validation set of 15 human labelled documents, there is room for improvement with larger validation sets. The semi-supervised LDA model confirmed our initial hypothesis that unsupervised approaches cannot easily replicate human extractive summarization and supervised approaches require large domain specific training sets which were not feasible for our timeframe. Through the process of modeling human text extraction, the importance of sentence embeddings was underscored through its usage during preprocessing stages to eliminate boilerplate text and during modeling in its use for unsupervised models.

Currently, no methods exist that can replicate a human text selection for 10-Q summarization, and our approach requires more improvement before being able to be deployed for widespread use.

Future Work

For future research, we would like to create a larger validation set to allow for a more robust semi-supervised approach to be taken. In addition, creating a front end for users to easily select companies, years, and summary lengths would greatly improve the practical use of our model. Given the current scores in precision, F1, and recall, reconsideration of current model selection would also be necessary once a larger validation set was created.

References

[1] G. S. L. Vishal Gupta, A Survey of Text Summarization Extractive Techniques, Chandigarh: Journal of Emerging Technologies in Web Intelligence, 2010.

[2] K. Wong, M. Wu and W. Li, "Extractive Summarization Using Supervised and Semi-supervised," *Proceedings of the 22nd International Conference on Computational Linguistics,* 2008.

[3] D. T. Araci, FinBert: Financial Sentiment Analysis with Pre-trained Language Models, Amsterdam: University of Amsterdam , 2019.

[4] alions7000, "SEC-EDGAR-TEXT," March 2020. [Online]. Available: https://github.com/alions7000/SEC-EDGAR-text.

[5] A. C. P. K. K. Nelson, "Litigation Risk and Voluntary Disclosure: The Use of Meaningful Cautionary Language," *SSRN Electronic Journal,* 2007.

[6] R. Michalcea and P. Tarau, "TextRank: Bringing Order into Texts," *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing,* 2004.

[7] D. Blei, A. Ng and M. Jordan, "Latent Dirichlet Allocation," *Advances in neural information,* 2002.

[8] J. Yeh, H. Ke, W. Yang and I. Meng, "Text summarization using a trainable summarizer and latent semantic analysis," *Information Processing & Management,* 2005.

[9] S. Xie, H. Lin and Y. Liu, "Semi-Supervised Extractive Speech Summarization," *INTERSPEECH 2010,* 2010.

[10] G. L'Huillier, A. Hevia, R. Weber and S. Rios, "Latent Semantic Analysis and Keyword Extraction," *Proceedings of the IEEE International Conference on Intelligence and Security Informatics,* 2010.

[11] A. Vashisht, "KL Sum algorithm for text summarization," OpenGenius IQ: Learn Computer Science, 2020. [Online].

Appendix A

Intra-Company 10-Q Sentence Similarity

We wanted to explore the similarity between 10-Qs created by the same company across multiple quarters to determine any structural similarities within a company's documents. We began by generating embeddings (or fixed-length vectors) for the sentences from the 10-Qs of a particular company. This process was done by the Universal Sentence Encoder (detailed in the *Preprocessing* section).

For the following plot, we concatenated the embeddings from Apple's three 10-Qs released during 2018 into a single list. We performed pairwise cosine similarity of every element in this list with every other element in this list, resulting in the cosine similarity heatmap in Figure 10. This cosine similarity heatmap simply takes the scores in the cosine similarity matrix, normalizes them, and replots them into a heatmap.

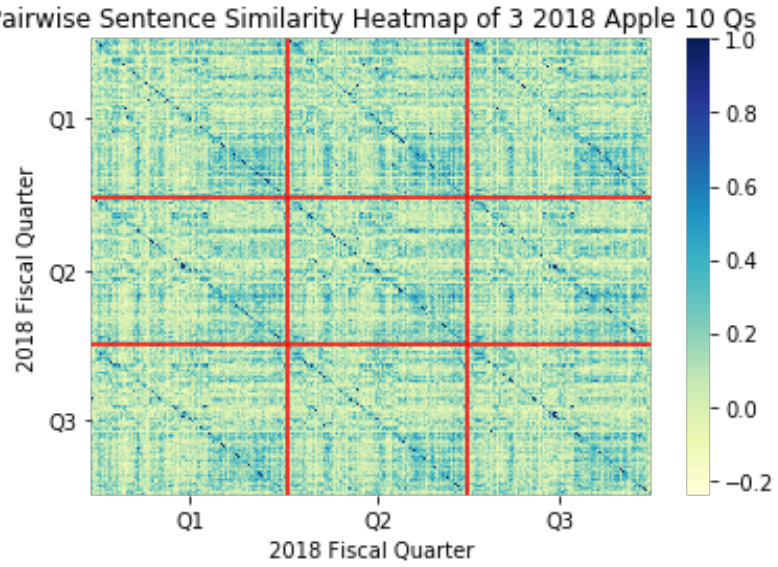


Figure 10: Pairwise Sentence Similarity Heatmap of 3 2018 Apple 10-Qs

The main diagonal is predictably pronounced since that depicts the similarity score of a sentence compared to itself, which should be 1. More interestingly, however, is the high degree of similarity along the off diagonals, which suggests that documents produced by the same company have significant structural similarity. In other words, a 10-Q produced by Apple may be remarkably similar to another 10-Q also produced by Apple on a sentence-by-sentence basis. Furthermore, 10-Qs produced within the same year by the same company could possibly have many sentences that are simply copied and pasted from one 10-Q to the next, only with small variation to reflect recent updates in the new quarter.

To ensure that this correlation was not specific to Apple, we compared the 10-Qs produced by other companies such as Oracle to each other, as seen in Figure 11.

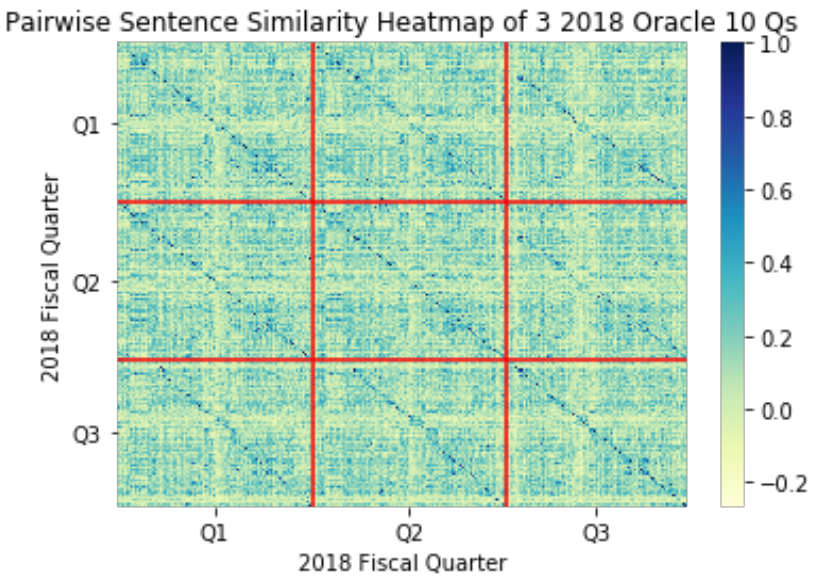


Figure 11: Pairwise Sentence Similarity of 3 2018 Oracle 10-Qs

Again, we see that there is a high degree of similarity along the off diagonals. This further reinforces our conclusion that there is high structural similarity among 10-Qs produced by the same company. However, when we plotted the 10-Qs of three different companies against each other, we saw no significant structural similarity between the documents in Figure 12.

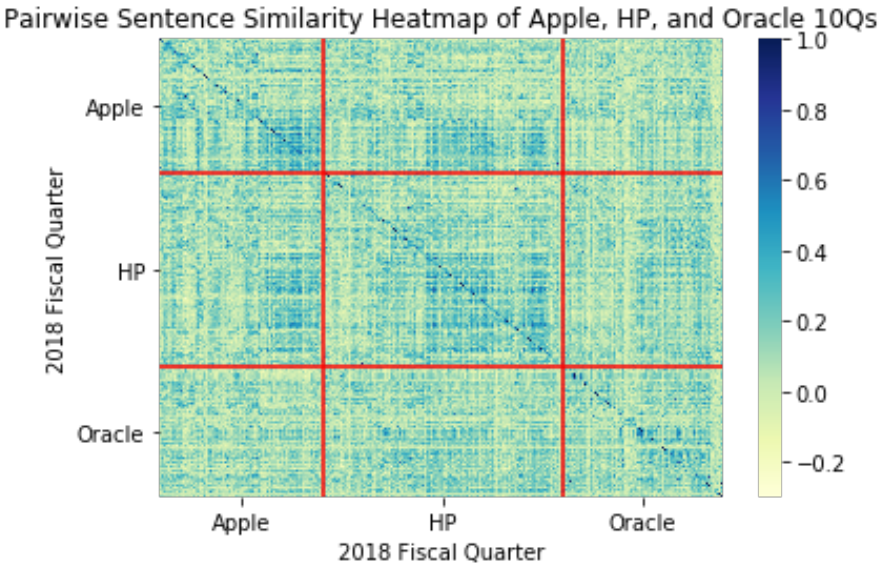


Figure 12: Pairwise Sentence Similarity of Apple, HP, and Oracle 10-Qs

This suggests that high structural similarity exists between documents produced by the same company. However, despite this high degree of structural similarity between documents from the same company, these similarity patterns did not give enough insight to determine whether we should remove those sentences or not. In other words, two highly similar sentences in similar positions of two consecutive 10-Qs did not indicate that the sentences were boilerplate or not.

Therefore, while these results were useful for data exploration to get a better understanding of the underlying patterns in the data, we ultimately did not pursue a preprocessing method or model from these results.

Appendix B

Other Models

In our modeling stage, we also tried the Kullback-Lieber (KL) Sum algorithm, Reduction, and Latent Semantic Analysis (LSA), which are all unsupervised approaches. Their performance results are shown in the figure below. We conclude that these models did not perform as well as the semi-supervised LDA model because they did not incorporate domain specific knowledge into their sentence selection processes.

f1_results

test filename	KL	LSA	Reduction
AAPL_0000320193_20180630.txt	0	0	0.07692307692307693
VSH_0000103730_20180331.txt	0	0.07692307692307693	0.15384615384615385
IBM_0000051143_20180331.txt	0	0	0.058823529411764705
AMAT_0000006951_20180729.txt	0.19354838709677422	0.12903225806451615	0.12903225806451615
CSCO_0000858877_20181027.txt	0	0.08888888888888889	0
MANT_0000892537_20180331.txt	0.24000000000000002	0.07142857142857144	0.14285714285714288
-	-	-	-
mean	0.07225806451612904	0.061045465884175566	0.09358036018377575
median	0.0	0.07417582417582419	0.10297766749379654
stdev	0.11290138247343154	0.051401294416917195	0.059261775948426666

Reduction:
Reduction is graph based summarization technique. Sentence importance is calculated by summing the weights of the edges of a sentence to other sentences. The weight of an edge between two sentences is calculated the same way that TextRank calculates weights. However, rather than using the page rank algorithm to determine node importance, Reduction simply sums the weights of the incoming edges.

Kullback-Lieber (KL) Sum:
KL Sum aims to minimize the summary vocabulary by checking the divergence from the input vocabulary. The method will greedily add sentences to a summary if it decreases the Kullback-Lieber divergence (relative entropy). The less this divergence, the more the summary and the document are like each other in terms of understandability and meaning (Vashisht, 2020).

Latent Semantic Analysis (LSA):
LSA is an unsupervised approach that uses a bag of words model to create a matrix of term occurrence for a document. The model assumes that words close in meaning will occur in similar pieces of text (which is especially relevant to our standardized text documents). An example matrix is shown below in Figure 13.

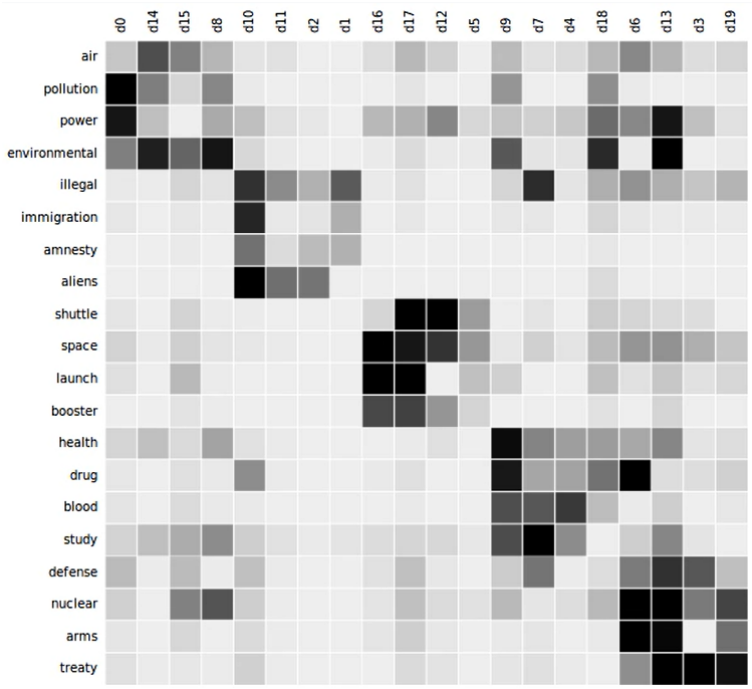


Figure 13: LSA matrix where columns are documents and rows are unique words. The darker a cell, the higher the weight of a word in a document.

LSA uses SVD on the matrix to determine the latent topics in the document. Text relationship map construction makes a mapping based on the semantic sentence representation from the semantic matrix created with SVD, and sentence selection is done through a global bushy path (created from the targeted paragraphs in the document) [8]. A generic path of this methodology done by [8] is shown in Figure 13. One challenge of this model is determining the number of topics in the given corpus text. A possible solution to this would be incorporating a topic coherence measure; a high value of topic coherence score model would be considered a good topic model.