Financial Statement Summarization with Transformers

Omkar Kharkar
(kharkaro@outlook.com)

Northwestern University

June 1, 2021

Abstract

Automatic financial statement summarization using machine learning is an important task for forecasting company performance. This paper tested 4 transformer architectures to determine their effectiveness for summarizing financial statements. Further applications of the methods presented here could include summarization of a wider variety of corporate documents, particularly for the private equity and venture capital markets.

Keywords— Transformers, Text Summarization, BERT, T5, PEGASUS, Financial Statements

Introduction

Financial statement analysis is an important part of any financial analyst's role. In finance, reading financial text data is particularly valuable for analysts to understand company earnings and determine equity market returns. Given the large volume of text data produced quarterly by public companies reporting earnings, there has been growing interest in understanding how automatic text summarization can help analysts to capture salient information in documents.

Text summarization is a task in the field of Natural Language Processing (NLP), which consists of utilizing machine learning algorithms to capture the salient parts of a document. Two variations of text summarization exist: Abstractive Summarization and Extractive Summarization. In Abstractive summarization, models generate a summary using some of the text used in the original document, but also include terminology that may not have appeared in the source text. In Extractive summarization, models capture the main ideas of the text, specifically by only using words appearing in the original document.

This paper examines several state-of-the-art transformers that have been used for abstractive text summarization. Specifically, this paper compares various fine-tuned transformer architectures to examine which models yield the best summarization performance, based on ROUGE scores (Lin 2004) for financial statements. While various pre-trained transformer models are currently available for text summarization, not all of them have been trained for the specific task of summarizing financial documents. Therefore, this work attempts to further examine how fine-tuned transformer models can handle terminology that is specific to the financial services domain.

Literature Review

Recent advancements in deep learning, particularly the transformer architecture, have greatly advanced NLP capabilities for all tasks, especially Text Summarization. Pytorch has emerged as the defacto library of choice for researchers, particularly in the NLP space. Dop 2020 provides a detailed and applied framework, that describes various neural architectures and examples of implementations in Pytorch (Dop 2020).

Vaswani et al. 2017 described in their paper the fundamentals of the attention mechanism, which is the building block of the transformer architecture. The authors discussed the merits of the transformer architecture, with the primary advantages being the training parallelization across GPUs in comparison to Recurrent Neural Networks (RNNs), and improved model performance due to the contextual relationships learned via the attention mechanism (Vaswani et al. 2017). This parallelization facilitates increasingly large language models, which has been a major contributing factor to the improved performance of transformers.

One of the early advancements in the field of transformers was the work of Peters et al. 2018 in developing the Embeddings from Language Models (ELMo) model. ELMo utilizes bi-directional LSTMs with attention layers to capture context-dependent word embeddings. This was a vast improvement over context-free word vectorization methods. (Peters et al. 2018).

Devlin et al. 2018 developed the idea behind Bidirectional Encoder Representations from Transformers (BERT). This utilizes the concept of bidirectional representations of text by analyzing the surrounding text in both the left and right directions to develop a deeper contextual understanding. It is trained in an unsupervised manner, allowing it to be trained on large datasets. BERT has achieved state-of-the-art performance on several NLP tasks, and has spawned numerous variants that have improved the benchmark performance in tasks such as question answering and text summarization. One of the characteristics of BERT is the use of masking, where some tokens are hidden, and the model is asked to fill in the masked tokens in the sentence (Devlin et al. 2018).

Financial data has recently been used for pre-training transformers in numerous studies to improve their predictive performance. For example, there are three different models, all named FinBERT, that were trained on different financial datasets. Araci 2019 developed one FinBERT variant, which was a BERT model enhanced with finance terminology for sentiment analysis. This was achieved by using several financial datasets. First, TRC-2, a Reuters dataset consisting of financial news articles, was filtered to emphasize financial terms during pre-training. Second, Financial PhraseBank was utilized for sentiment analysis. This dataset consisted of 4,845 English sentences from LexisNexis, pertaining to financial news stories and annotated by human reviewers (Araci 2019). The result was that this version of FinBERT was able to outperform the other benchmark methods, including transformers in detecting sentiment due varied financial terminology used.

Desola, Hanna, and Nonis 2019 trained another variant called FinBERT Prime by directly using SEC 10-K filings for various US companies. FinBERT Prime differentiated itself by being a from-scratch model; it was trained purely for financial documents, without any pre-training on the general English corpus, as commonly used by the base BERT model. FinBERT Prime and the tested variants were found to outperform BERT in next-sentence predictions (Desola, Hanna, and Nonis 2019).

In contrast to the techniques utilized by Desola, Hanna, and Nonis 2019, another FinBERT by Liu 2019a was pre-trained in the traditional manner, utilizing both general English corpora such as Wikipedia and BooksCorpus while being supplemented by finance-specific datasets such as FinancialWeb and YahooFinance. Liu et al. (2020) tested their FinBERT model and found performance improvements in comparison to BERT for Financial Question Answering, Sentiment Analysis and Sentence Boundary Detection tasks (Liu 2019a). However, it should be noted that this is an unorthodox step, since most BERT variants have already been trained on generic English corpora, and typically require fine-tuning using a domain-specific corpus.

Text summarization as an NLP task has also benefitted from the use of Transformer models. Zhang et al. 2019 developed the Pre-training with Extracted Gap-sentences for Abstractive Summarization (PEGASUS). PEGASUS has achieved state-of-the-art performance in text summarization. The main differentiation in comparison to other architectures is the masking of sentences from input documents which are generated together as an output sequence, known as Gap Sentences Generation (GSG). GSG, coupled with the masking commonly found in BERT, has greatly improved performance across a variety of English corpora (Zhang et al. 2019). Passali et al. 2021 introduced a version of PEGASUS that was fine-tuned specifically the Bloomberg Market and Financial News API, demonstrating strong performance on financial news summarization. This model is versatile, having been trained documents ranging from 20 - 3,578 words. This resulted in a significantly higher ROUGE score, in comparison to the baseline PEGASUS model trained solely on the XSUM dataset Passali et al. 2021.

Lewis et al. 2019 introduced the BART architecture which utilizes a denoising autoencoder and sequence-to-sequence models, as part of the Transformer architecture. In contrast to several of the architectures discussed in this section, BART is trained by corrupting the input text using a noising function, and then rebuilding the original text using a transformer model (Lewis et al. 2019). While the original transformer architecture (Vaswani et al. 2017) is retained, BART makes several modifications to BERT by not utilizing the feed-forward network structure commonly used in BERT models (Lewis et al. 2019).

Liu 2019b introduced the idea of BERTSUM, a variation of BERT that performs extractive text summarization (summaries using words contained in the text), as opposed abstractive text generation (generating a short summary that captures the important ideas), cited in the various research papers above. BERTSUM utilizes a pre-trained BERT model as the foundation but does not utilize a sigmoid classifier and instead applies additional transformer layers on the sentence representations. Additionally, BERTSUM utilizes Long-Short Term Memory (LSTM) layers over the BERT outputs to improve the summarization capabilities (Liu 2019b).

Qi et al. 2021 recently developed a new model, ProphetNet-X, which has demonstrated strong performance on both text summarization and question generation tasks. ProphetNet-X builds on the success of transformer models by adding future token predictions for sentences, a hallmark of the ProphetNet architecture on which the models are based. Fundamentally, the model uses transformer encoder-decoder layers to train on various n-grams encountered in any text used for pre-training. ProphetNet has demonstrated state-of-the-art results on News Title Generation (NTG) and Question Generation (QG) tasks, demonstrating the continued versatility of transformer network (Qi et al. 2021).

Raffel et al. 2019 developed the text-to-text transformer (T5) architecture which has been used for a variety of NLP tasks. The model was developed using the attention mechanism, but has a slightly different structure, utilizing a encoder-decoder architecture in order to take text as an input, and output text. Additionally, this model was trained on both supervised and unsupervised learning tasks. T5 has demonstrated strong performance on summarization and question answering tasks, and was one of the candidate models tested in this study (Raffel et al. 2019)

Rothe, Narayan, and Severyn 2019 developed BERT2BERT, a variant of the

BERT architecture referenced above that utilizes an encoder-decoder architecture with both the encoder and decoder layers obtained from pre-trained BERT layers. The main benefit of this model is that does not have the same size as BERT pre-trained layers, but still achieves good performance on abstractive summarization tasks (Rothe, Narayan, and Severyn 2019).

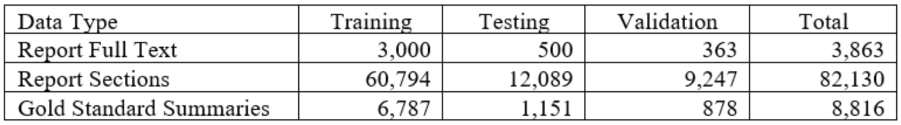
Lundberg and Lee 2017 developed the SHAP library, a Python library for machine learning interpretability. The SHAP library calculates Shapley Additive Explanations for features used in a machine learning model. One of the key strengths of the library is its applicability to numeric, categorical, text and image data. This allows users to gain insights into models, and explainability into what factors drove a model's predictions. Recently, the SHAP library was updated to incorporate predictions for transformer models (Lundberg and Lee 2017). Model interpretability is further discussed in the sections below.

Data

The data was obtained from the MultiLing 2019 Financial Narrative Summarization task (Giannakopoulos 2019). MultiLing 2019 was a workshop that was hosted as part of the Recent Advances in Natural Language Processing (RANLP) 2019 conference. The Financial Narrative Summarization task was a challenge to measure the effectiveness of summaries produced by machine learning models. The data comprises of data obtained from UK firms listed on the London Stock Exchange (LSE) (El-Haj 2019).

Table 1 below shows the division of the dataset into training, validation and testing components. Since the reports are varied across companies, each report is further divided into sections, such as Financial Highlights, Notes to Financial Statements, Auditor's Reports, etc. Not all sections from the full report have been considered for summarization; sections consisting entirely of financial statement ledgers, which are not easily summarized using English text, have been omitted from evaluation. The gold-standard summaries are human-generated summaries for the Chairman's CEO's statements and the Highlights sections that can be used as the basis of comparison against model-generated summaries for the corresponding section (El-Haj 2019).

Table 1: Original Training Data for Financial Narrative Summarization



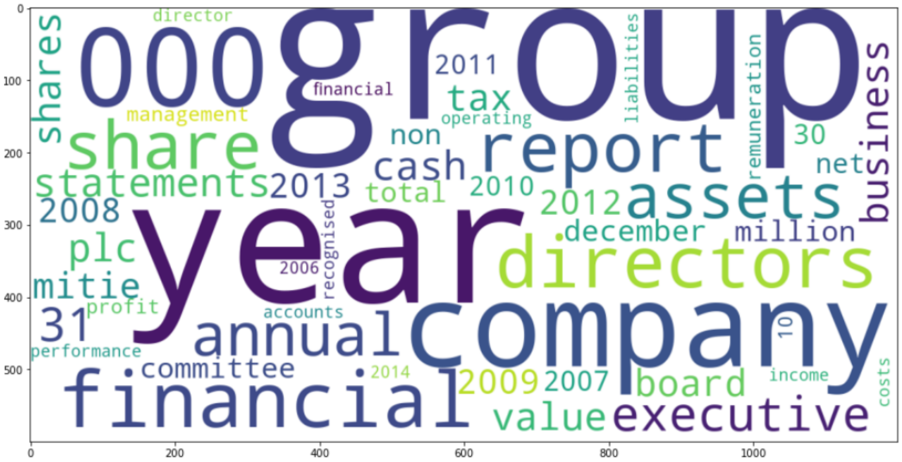
Data Type	Training	Testing	Validation	Total
Report Full Text	3,000	500	363	3,863
Report Sections	60,794	12,089	9,247	82,130
Gold Standard Summaries	6,787	1,151	878	8,816

Each original report contained two gold summaries, corresponding to either the Highlights section or the Chairman's/CEO statement (management commentary) sections. This study focused on the summarization of the management commentary sections, for each report. Given the range of text lengths present in the dataset, the analysis was restricted to sections containing between 1,100 - 2,500 words, which constituted the middle 50% of report lengths. For each gold standard summary, the first

10% of the summary length was used as the baseline of comparison against model-generated summaries. This was done mainly due to the transformers utilized, and the lengths of the summaries generated. This resulted in a dataset of 1,528 summaries total. The data was further divided into a 80/20 split for training and validation respectively, corresponding to 1,222 summaries used for training, and 306 summaries used for validation.

The full-text reports were initially processed and cleaned to remove text, and understand the overall themes observed in the reports. Examples of text that was removed included hyperlinks and non-ASCII characters that could be read in from the original text. Figure 1 below shows the initial results for the term-frequency inverse document frequency (TF-IDF) word clouds that were generated to understand that most frequent n-grams in the documents.

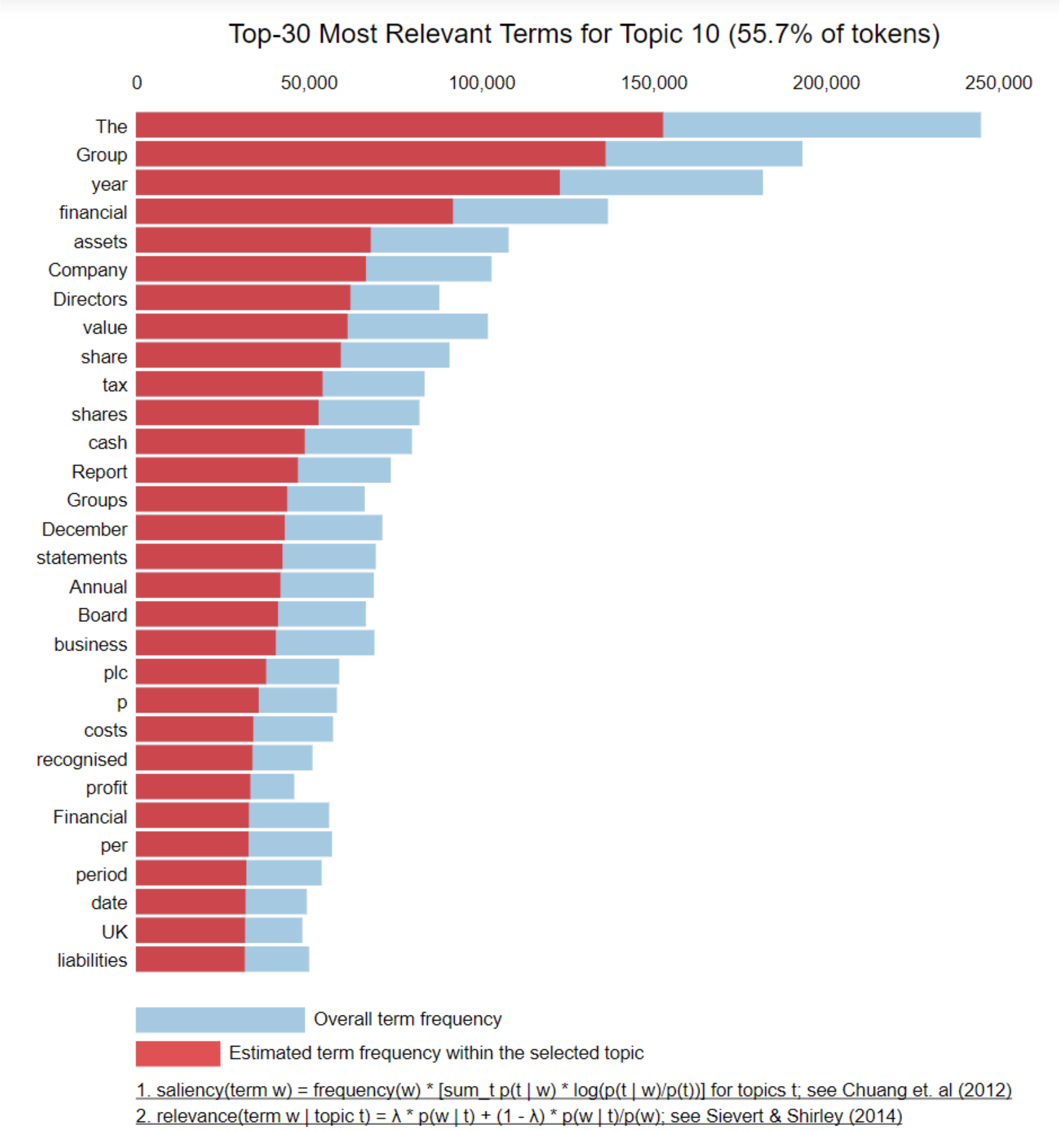
Figure 1: TF-IDF Word Clouds



The TF-IDF word clouds shown above were extremely useful for exploratory data analysis by understanding common terms across documents, to clean the data.

Latent Dirichlet Allocation (LDA) was utilized in order to generate topics that could describe the set of documents, as well diagnose any issues related to data cleaning, as seen in Figure 2 below.

Figure 2: Latent Dirichlet Allocation (LDA) Topic Models for Full Reports



Top-30 Most Relevant Terms for Topic 10 (55.7% of tokens)

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

LDA topic models were useful for understand the various groupings of words and discover any characters that needed to be cleaned (removed) from the text. Here, the top 30 words that form topic 10, one of the largest topics (themes) prevalent throughout the dataset, is shown here. This topic most likely reflects management's commentary on financial performance.

Methods

There were several different methods that were used in order to draw insights from the data. First, in order to perform exploratory analysis on the data, term-frequency inverse-document frequency (TF-IDF) word clouds were constructed to understand the most frequent phrases that appeared in the documents, along with Latent Dirichlet Allocation (LDA), as described above.

For text summarization, 4 transformer model architectures were fine-tuned on the financial statements original text, along with the gold standard summaries (as defined above) to compare the summarization results. These model architectures were T5, BERT2BERT, PEGASUS-Finance and PEGASUS-Legal. PEGASUS-Legal is a transformer that is similar to the PEGASUS architecture described above, but has been further fine-tuned on 2,700 litigation documents from the Securities and Exchange Commission (SEC) (Sairam 2021). The T5, BERT2BERT and PEGASUS-Finance were each trained for 2 and 5 epochs, with an input of 512 tokens and output of 150 tokens (the maximum possible output length for these models). Due to computational constraints, PEGASUS-Legal could not be trained for more than 2 epochs, but instead, different output summary lengths (150 and 250 tokens) were used for comparing results.

Results

The ROUGE-1 and ROUGE-L metrics (Lin 2004) were used for evaluating the quality of the model generated summaries for the validation dataset. ROUGE-1 measures the overlap of the n-grams (in this case n = 1) between the gold standard summary and the model generated summary.

$$ROUGE_N = \frac{\Sigma Count_{matching}}{\Sigma Count_{total}}$$

This formula gives rise to 2 variations:

$$ROUGE_{NPrecision} = \frac{\Sigma Count_{matching}}{\Sigma Count_{ModelSummary}}$$

Precision measures the quality of words generated in the summary, by looking at how many of the words in the generated summary were relevant.

$$ROUGE_{NRecall} = \frac{\Sigma Count_{matching}}{\Sigma Count_{GoldSummary}}$$

Recall measures the quantity of the words captured by the summary, by comparing it against the number of words in the gold standard summary.

Together, the harmonic mean of Precision and Recall can be used to compute the F-measure, an overall measure of a model's performance, where β is a measure of weighting given to precision over recall, depending on which metric may be of more importance.

$$F_{measure} = \frac{(1+\beta^2)(Recall)(Precision)}{Recall + (\beta^2 Precision)}$$

ROUGE-L measures the longest common subsequence (LCS) of matching n-grams between the model generated text and gold standard summary text. This allows for the evaluation of a model summary using sentence-level word order, as opposed to only looking at n-grams. Similarly, precision and recall for ROUGE-L can also be computed (Lin 2004):

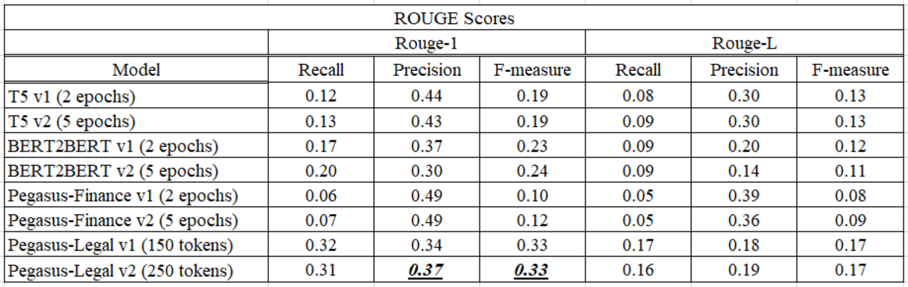
$$ROUGE_L = \frac{LCS_{matching}}{SummaryLength}$$

$$ROUGE_{LPrecision} = \frac{LCS_{matching}}{ModelSummaryLength}$$

$$ROUGE_{LRecall} = \frac{LCS_{matching}}{GoldSummaryLength}$$

The mean ROUGE-1 and ROUGE-L scores for each of the transformer models analyzed in this study computed on the validation dataset are shown in the Table below:

Table 2: ROUGE Scores for Fine-Tuned Transformers



ROUGE Scores						
	Rouge-1			Rouge-L		
Model	Recall	Precision	F-measure	Recall	Precision	F-measure
T5 v1 (2 epochs)	0.12	0.44	0.19	0.08	0.30	0.13
T5 v2 (5 epochs)	0.13	0.43	0.19	0.09	0.30	0.13
BERT2BERT v1 (2 epochs)	0.17	0.37	0.23	0.09	0.20	0.12
BERT2BERT v2 (5 epochs)	0.20	0.30	0.24	0.09	0.14	0.11
Pegasus-Finance v1 (2 epochs)	0.06	0.49	0.10	0.05	0.39	0.08
Pegasus-Finance v2 (5 epochs)	0.07	0.49	0.12	0.05	0.36	0.09
Pegasus-Legal v1 (150 tokens)	0.32	0.34	0.33	0.17	0.18	0.17
Pegasus-Legal v2 (250 tokens)	0.31	***0.37***	***0.33***	0.16	0.19	0.17

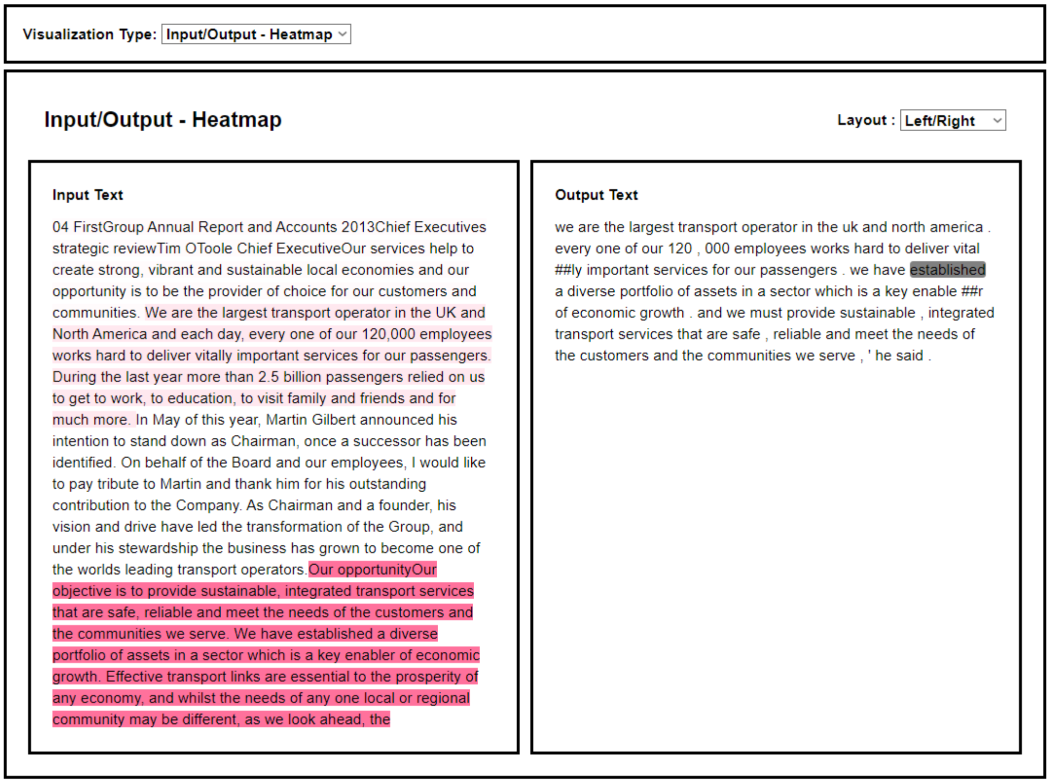
Based on the ROUGE-1 and ROUGE-L scores obtained for the validation set, PEGASUS-Legal v2 was the top performing model, with the highest Rouge-1 F-measure and Precision scores, followed by PEGASUS-Legal v1 and BERT2BERT v2.

Analysis & Interpretation

Based on the results obtained above, PEGASUS-Legal v2 was statistically the best model. For all models, the use of SHAP values, particularly Partition Explainer, provides interpretability into which words were important to a particular model-generated summary.

Figures 3 - 5 below show an example of SHAP values generated for one of the reports, for some of the model outputs:

Figure 3: BERT2BERT v1 Summary SHAP Values



Visualization Type: Input/Output - Heatmap

Input/Output - Heatmap

Layout : Left/Right

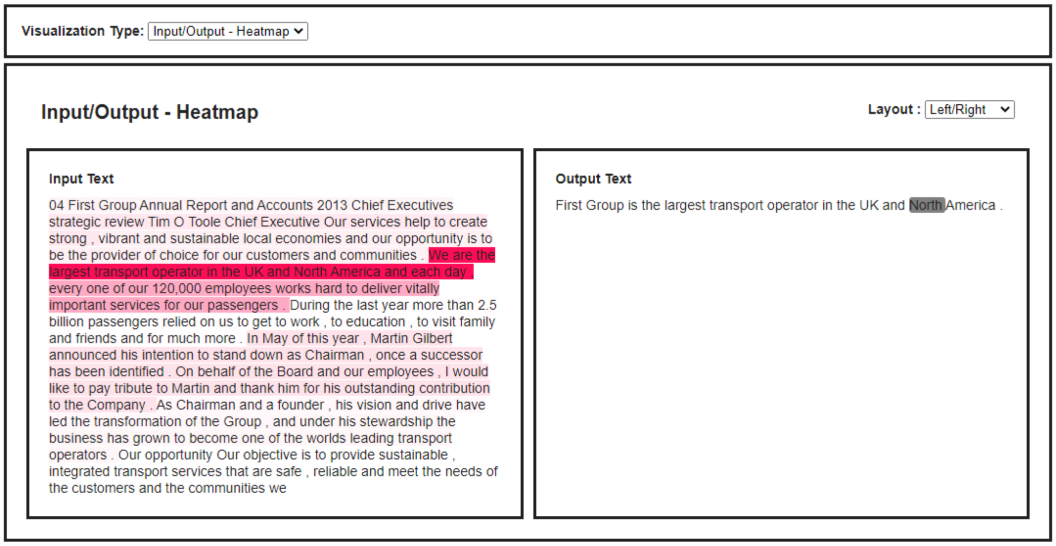
Input Text

04 FirstGroup Annual Report and Accounts 2013Chief Executives strategic reviewTim OToole Chief ExecutiveOur services help to create strong, vibrant and sustainable local economies and our opportunity is to be the provider of choice for our customers and communities. We are the largest transport operator in the UK and North America and each day, every one of our 120,000 employees works hard to deliver vitally important services for our passengers. During the last year more than 2.5 billion passengers relied on us to get to work, to education, to visit family and friends and for much more. In May of this year, Martin Gilbert announced his intention to stand down as Chairman, once a successor has been identified. On behalf of the Board and our employees, I would like to pay tribute to Martin and thank him for his outstanding contribution to the Company. As Chairman and a founder, his vision and drive have led the transformation of the Group, and under his stewardship the business has grown to become one of the worlds leading transport operators.Our opportunityOur objective is to provide sustainable, integrated transport services that are safe, reliable and meet the needs of the customers and the communities we serve. We have established a diverse portfolio of assets in a sector which is a key enabler of economic growth. Effective transport links are essential to the prosperity of any economy, and whilst the needs of any one local or regional community may be different, as we look ahead, the

Output Text

we are the largest transport operator in the uk and north america . every one of our 120 , 000 employees works hard to deliver vital ##ly important services for our passengers . we have established a diverse portfolio of assets in a sector which is a key enable ##r of economic growth . and we must provide sustainable , integrated transport services that are safe , reliable and meet the needs of the customers and the communities we serve , ' he said .

Figure 4: PEGASUS-Finance v2 Summary SHAP Values



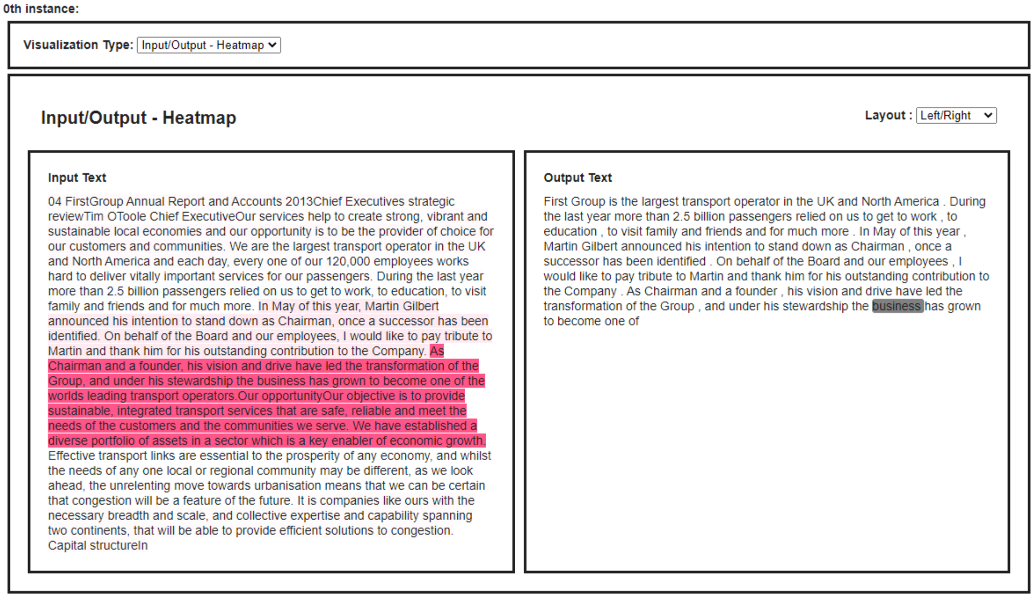
Input Text

04 First Group Annual Report and Accounts 2013 Chief Executives strategic review Tim O Toole Chief Executive Our services help to create strong , vibrant and sustainable local economies and our opportunity is to be the provider of choice for our customers and communities . We are the largest transport operator in the UK and North America and each day , every one of our 120,000 employees works hard to deliver vitally important services for our passengers . During the last year more than 2.5 billion passengers relied on us to get to work , to education , to visit family and friends and for much more . In May of this year , Martin Gilbert announced his intention to stand down as Chairman , once a successor has been identified . On behalf of the Board and our employees , I would like to pay tribute to Martin and thank him for his outstanding contribution to the Company . As Chairman and a founder , his vision and drive have led the transformation of the Group , and under his stewardship the business has grown to become one of the worlds leading transport operators . Our opportunity Our objective is to provide sustainable , integrated transport services that are safe , reliable and meet the needs of the customers and the communities we

Output Text

First Group is the largest transport operator in the UK and North America .

Figure 5: PEGASUS-Legal v2 Summary SHAP Values



Input Text

04 FirstGroup Annual Report and Accounts 2013Chief Executives strategic reviewTim OToole Chief ExecutiveOur services help to create strong, vibrant and sustainable local economies and our opportunity is to be the provider of choice for our customers and communities. We are the largest transport operator in the UK and North America and each day, every one of our 120,000 employees works hard to deliver vitally important services for our passengers. During the last year more than 2.5 billion passengers relied on us to get to work, to education, to visit family and friends and for much more. In May of this year, Martin Gilbert announced his intention to stand down as Chairman, once a successor has been identified. On behalf of the Board and our employees, I would like to pay tribute to Martin and thank him for his outstanding contribution to the Company. As Chairman and a founder, his vision and drive have led the transformation of the Group, and under his stewardship the business has grown to become one of the worlds leading transport operators.Our opportunityOur objective is to provide sustainable, integrated transport services that are safe, reliable and meet the needs of the customers and the communities we serve. We have established a diverse portfolio of assets in a sector which is a key enabler of economic growth. Effective transport links are essential to the prosperity of any economy, and whilst the needs of any one local or regional community may be different, as we look ahead, the unrelenting move towards urbanisation means that we can be certain that congestion will be a feature of the future. It is companies like ours with the necessary breadth and scale, and collective expertise and capability spanning two continents, that will be able to provide efficient solutions to congestion. Capital structureIn

Output Text

First Group is the largest transport operator in the UK and North America . During the last year more than 2.5 billion passengers relied on us to get to work , to education , to visit family and friends and for much more . In May of this year , Martin Gilbert announced his intention to stand down as Chairman , once a successor has been identified . On behalf of the Board and our employees , I would like to pay tribute to Martin and thank him for his outstanding contribution to the Company . As Chairman and a founder , his vision and drive have led the transformation of the Group , and under his stewardship the business has grown to become one of

Based on the SHAP values for generated summaries, PEGASUS-Legal v2 generates longer summaries while incorporating financial and legal terminology, resulting in larger ROUGE scores. However, this performance comes at the cost of computational time. T5, BERT2BERT and PEGASUS-Finance required a training time of approx. 5 minutes per epoch, while PEGASUS-Legal required approx. 15 minutes per epoch. One challenge with the BERT2BERT models was the repetition of certain words. Tuning grid-search parameters, such as beam search and increasing the n-gram penalty parameters could correct this problem. The PEGASUS-Finance models both achieved a high ROUGE-1 precision of 0.49 and ROUGE-L precisions of 0.39 and 0.36, respectively. This is consistent with the fact that the PEGASUS-Finance models were pre-trained using articles from Bloomberg on finance topics, signifying that the models were able to incorporate finance vocabulary in the summaries. However, T5, BERT2BERT and PEGASUS-Finance summary lengths are very short, resulting in lower recall metrics.

All work was done on using a GTX 1080 Ti (11 GB RAM) and a Tesla T4 (16 GB RAM) GPU. This played a major factor in the choice of transformer architectures, size of the input text and output summary lengths. The main challenge of this study was the large size of transformers. Many models such as the base BERT and base PEGASUS models are large (340 M and 568 M parameters, respectively), and require extensive GPU memory for computation. This influenced factors such as the batch size and the maximum length of input tokens that could be used. For the 2 PEGASUS models, an input of length 512 tokens was used, since the maximum input length of 1,024 tokens was infeasible given GPU memory limitations.

Conclusions

Based on the results obtained for financial statement summarization, the PEGASUS-Legal v2 model is the optimal model with the highest ROUGE scores, with ROUGE-1 precision at 0.37 precision and an F-measure of 0.33. This model outperformed the T5, BERT2BERT and PEGASUS-Finance models for the summarization task.

Directions for Future Work

There are several further extensions to this work. First, one improvement to model performance would be fine-tune models on the A100/V100 data center GPUs, which offer more processing power and memory. This would allow for the use of larger input sequences, batch sizes, and larger output summaries.

Second, alternative transformer architectures could be compared against the models used here, particularly ones that are designed for handling larger input text lengths. Examples include the BIGBIRD PEGASUS and Longformer architectures, which have demonstrated success for summarization on long sequences, as commonly found in financial statements. BIGBIRD utilizes a sparse attention mechanism, along with maximum length of 4,096 tokens, and demonstrated considerable improvements in ROUGE score for summarization, in comparison to RoBERTa and PEGASUS (Zaheer et al. 2020). The Longformer architecture uses attention on both local and global scales, allowing only half the text to be scanned at any point in time. This idea is similar to the sliding window concept used in Convolutional Neural Networks, and reduces the $O(n^2)$ complexity of the attention mechanism to an $O(n)$ complexity, allowing for longer input text lengths (up to 4,096 tokens) to be utilized by the model (Beltagy, Peters, and Cohan 2020).

Third, models that have been pre-trained on industry-specific texts demonstrated superior ROUGE precision scores, as seen by PEGASUS-Finance and PEGASUS-Legal. An extension of this study would be to apply these models to the private equity and venture capital industries, where text summarization can play an important role in understanding private company valuation by analyzing internal corporate documents.

References

Araci, Dogu. 2019. "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models." *CoRR* abs/1908.10063. arXiv: 1908.10063. http://arxiv.org/abs/1908.10063.

Beltagy, Iz, Matthew E. Peters, and Arman Cohan. 2020. "Longformer: The Long-Document Transformer." *CoRR* abs/2004.05150. arXiv: 2004.05150. https://arxiv.org/abs/2004.05150.

Desola, Vinicio, Kevin Hanna, and Pri Nonis. 2019. "FinBERT: pre-trained model on SEC filings for financial natural language tasks" (August). https://doi.org/10.13140/RG.2.2.19153.89442.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *CoRR* abs/1810.04805. arXiv: 1810.04805. http://arxiv.org/abs/1810.04805.

Dop, Thomas. 2020. *Hands-On Natural Language Processing with Pytorch 1.x.* Edited by David Sugarman. Birmingham, UK: Packt Publishing Ltd.

Giannakopoulos, George. 2019. "Task: Financial narrative summarization," July 10, 2019. http://multiling.iit.demokritos.gr/pages/view/1648/task-financial-narrative-summarization.

El-Haj, Mahmoud. 2019. "MultiLing 2019: Financial Narrative Summarisation." In *Proceedings of the Workshop MultiLing 2019: Summarization Across Languages, Genres and Sources,* 6–10. Varna, Bulgaria: INCOMA Ltd., September. http://doi.org/10.26615/978-954-452-058-8_002.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." *CoRR* abs/1910.13461. arXiv: 1910.13461. http://arxiv.org/abs/1910.13461.

Lin, Chin-Yew. 2004. “ROUGE: a Package for Automatic Evaluation of Summaries.” In *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain,* 74–81. July. https://www.microsoft.com/en-us/research/publication/rouge-a-package-for-automatic-evaluation-of-summaries/.

Liu, Yang. 2019a. “Fine-tune BERT for Extractive Summarization.” *CoRR* abs/1903.10318. arXiv: 1903.10318. http://arxiv.org/abs/1903.10318.

———. 2019b. “Fine-tune BERT for Extractive Summarization.” *CoRR* abs/1903.10318. arXiv: 1903.10318. http://arxiv.org/abs/1903.10318.

Lundberg, Scott M, and Su-In Lee. 2017. “A Unified Approach to Interpreting Model Predictions.” In *Advances in Neural Information Processing Systems 30,* edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4765–4774. Curran Associates, Inc. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Passali, Tatiana, Alexios Gidiotis, Efstathios Chatzikyriakidis, and Grigorios Tsoumakas. 2021. “Towards Human-Centered Summarization: A Case Study on Financial News.” In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing,* 21–27. Online: Association for Computational Linguistics, April. https://www.aclweb.org/anthology/2021.hcinlp-1.4.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. “Deep contextualized word representations.” *CoRR* abs/1802.05365. arXiv: 1802.05365. http://arxiv.org/abs/1802.05365.

Qi, Weizhen, Yeyun Gong, Yu Yan, Can Xu, Bolun Yao, Bartuer Zhou, Biao Cheng, et al. 2021. “ProphetNet-X: Large-Scale Pre-training Models for English, Chinese, Multi-lingual, Dialog, and Code Generation.” *CoRR* abs/2104.08006. arXiv: 2104.08006. https://arxiv.org/abs/2104.08006.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.” *CoRR* abs/1910.10683. arXiv: 1910.10683. http://arxiv.org/abs/1910.10683.

Rothe, Sascha, Shashi Narayan, and Aliaksei Severyn. 2019. “Leveraging Pre-trained Checkpoints for Sequence Generation Tasks.” *CoRR* abs/1907.12461. arXiv: 1907.12461. http://arxiv.org/abs/1907.12461.

Sairam, Naren. 2021. “PEGASUS for legal document summarization,” May 31, 2021. https://huggingface.co/nsi319/legal-pegasus.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” *CoRR* abs/1706.03762. arXiv: 1706.03762. http://arxiv.org/abs/1706.03762.

Zaheer, Manzil, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, et al. 2020. “Big Bird: Transformers for Longer Sequences.” *CoRR* abs/2007.14062. arXiv: 2007.14062. https://arxiv.org/abs/2007.14062.

Zhang, Jingqing, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization.” *CoRR* abs/1912.08777. arXiv: 1912.08777. http://arxiv.org/abs/1912.08777.