An Exploration of Automatic Text Summarization of Financial Reports

Samir Abdaljalil
Carnegie Mellon University in Qatar
sabdalja@alumni.cmu.edu

Houda Bouamor
Carnegie Mellon University in Qatar
hbouamor@andrew.cmu.edu

Abstract

In the domain of finance, documents tend to be long, averaging at approximately 180 pages. This creates a need for finding efficient ways to use technology to leverage the existence of these huge amounts of textual data. This goes hand in hand with the pressing need to make investment/financial decisions in a fast manner to ensure maximized financial gain. However, exhaustive reading of financial documents such as annual reports is extremely laborious. Hence, automatic summarization methods could simplify this task and provide access to a smaller but informative chunk of a given document. In this paper, we explore several approaches for summarizing the qualitative sections of annual reports using extractive summarization, Natural Language Processing (NLP), machine learning, and deep learning techniques. We investigate multiple approaches by defining two different tasks: a sentence-based summarization task and a section-based summarization one. The latter is tailored to the structure of the annual reports. We evaluate the quality of the summaries using an existing dataset of annual reports published by British firms belonging to the London Stock Exchange and their corresponding manually created summaries built for the 2020 FNS Shared task. Our best model makes use of the power of unsupervised clustering techniques to group sections based on their meaning and achieves a ROUGE-L score of 36%.

1. Introduction

As technological resources are evolving, different domains are starting to adopt technology, and make use of it in a way that makes certain aspects of information access in each domain more efficient. The digital world has become infiltrated with massive volumes of digital data. In 2018, the size of the indexed World Wide Web was over 5.22 billion pages (Kunder, 2018), spread over 1.8 billion websites (Fowler, 2018). Particularly, in the domain of finance, as the number of electronic text documents is growing, the need for automating several tasks leveraging them is increasing. Initially, there was a lack of technological resources (datasets and systems) enabling individuals in the field to perform their jobs efficiently. However, over the years the emergence of the Fintech industry disrupted the way individuals operate in the Finance domain. In fact, according to PricewaterhouseCoopers, "Funding of FinTech startups has increased at a compound annual growth rate (CAGR) of 41% over the last four years, with over US$40 billion in cumulative investment." (PWC, 2018) According to JP Morgan, the Banking and Securities industry has been widely investing in Artificial Intelligence (AI) applications, a good example being the usage of news sentiment analysis for automatic investment (2020).

This sudden surge of investment in the FinTech industry is understandable, as from this domain comes a large amount of data created by different firms. Such data comes in different forms like annual financial reports, quarterly reports, preliminary earnings announcements, conference calls and press releases (El-Haj *et al.*, 2018). However, most of these documents are published in a PDF format including figures, tables, numbers, and most importantly textual narratives. Also, financial documents can be quite long, averaging at approximately 180 pages (Leidner, 2019). This creates a need for finding efficient ways that make use of technology to leverage the existence of these textual datasets, especially to extract the most relevant information from different key sections. This is where automatic text summarization systems come into play, as individuals in the field would be able to go through summaries of reports and other related documents, and derive appropriate market conclusions (Khant & Singh Mehta, 2018).

With the rise of the application of AI in automating certain processes, the use of NLP piques the interest of professionals in the Finance domain, since these financial reports largely consist of textual data. A substantial amount of research has been conducted on the use of NLP and text-mining techniques in the financial domain by looking into sentiment analysis and information extraction techniques applied to Financial news (Filippova *et al.*, 2009). Despite the several efforts (El-Haj *et al.,* 2019, El-Haj *et al.*, 2020) and the importance of the task, research on how to use NLP, machine learning, and deep learning techniques to analyze and summarize other textual datasets in the finance domain like Annual reports has not seen major development yet, due to the extremely unstructured nature of those reports and to the scarcity of annotated datasets.

There are two general approaches to automatic summarization: extractive and abstractive. In this research, we will be exploring extractive summarization approaches

Proceedings of the Third Workshop on Financial Technology and Natural Language Processing
(FinNLP@IJCAI 2021), pages 1-7, Online, August 19, 2021.

due to the nature of the datasets we use in our experiments. To the best of our knowledge, there is no dataset of annual reports and their equivalent abstractive summaries publicly available. Building such a dataset is expensive in terms of time and funding, as it requires expertise in the domain, beyond any linguistic knowledge. Extractive summarization methods consist in first identifying the important paragraphs or sentences from a given input document. Then, in general a ranking function is applied to keep only the most informative and important subset to include in the final summary (Gupta & Lehal, 2010). In this work, we explore the use of different approaches and build several automatic financial narrative summarization models. We evaluate the quality of the models under different experimental setups. Our best model is the one in which we combine pretrained language models' robustness in textual representation with the power of unsupervised clustering techniques into grouping similar sentences based on their meaning and not on a surface word-level. Our best model achieves a ROUGE-L of 36%.

2. Related Work

In the domain of finance, annual reports have been heavily studied in several research works. For instance, Stanton & Stanton (2002), discuss the importance of a financial annual report to a firm's activities and define the role it plays. Similarly, Ghazali & Annum (2010) look at the usefulness of having a corporate annual report for companies in Malaysia and express the importance of an annual report and its effect on any company's image. Annual reports allow prospective and/or current customers to build a sense of confidence towards a brand as it offers a clear view of any changes in operation, and the profitability profile of any company.

Automatic extractive summarization has been studied in several research works. Leidner (2019) explores the different NLP techniques that are used in text summarization. He discusses the typical quality dimensions of a financial summary and the different methods used in summarizing financial documents by looking into heuristic, statistical, and neural models. Abujar *et al.* (2017) propose a heuristic approach to extractive summarization of Bengali text by identifying tokenized word frequency scores, and deducing sentence scores, which would identify the most important sentences to be included in the summary. Graph-based methods have also been widely explored. Xu *et al.* (2013) propose a graph-based model for multi-tweet extractive summarization. Their model leverages the Named Entities, and frequency of topics discussed within the tweets. Similarly, Mihalcea (2004) applies the graph-based TextRank algorithm to single-document summarization of news articles, which ranks sentences based on their connections and similarity scores between other sentences within the news article. Query-based automatic summarization has also been explored. Fillippova *et al.* (2009), presents an extractive summarization system for summarizing financial news. Their model takes a company's name as input and retrieves any financial news regarding that company posted on Yahoo News. Then they rank sentences in terms of importance and relevance. Furthermore, Berger & Mittal (2000) employ a statistical approach to query-based extractive summarization, by making use of frequently asked questions documents found on websites where each answer in a FAQ is considered as a summary of the document relative to the question which preceded it. Other researchers defined the task of extractive summarization as a typical classification task. Chuang & Yang (2000) presented a system for U.S Patent and Trademark documents summarization. They defined a total of 23 features and explored several machine learning algorithms, including DistAI.

In the domain of finance, recently, statistical features with heuristic approaches have been used to summarize financial disclosure texts (Cardinaels *et al.*, 2019), generating summaries with reduced positive bias and leading to more conservative valuation judgements by investors that receive them. Furthermore, the financial narrative summarization task (El-Haj *et al.*, 2019) of the Multiling 2019 workshop (Giannakopoulos, 2019) involved the generation of structured summaries from financial narrative disclosures. It aimed to provide researchers in the field of NLP with a platform to explore the different approaches of extractive automatic summarization to UK annual reports, while also demonstrating the value and challenges of applying automatic text summarization to financial text written in English, usually referred to as financial narrative disclosures (El-Haj *et al.*, 2019). This task was extended to create the FNS 2020 Shared Task (El-Haj *et al.*, 2020) co-located with the 2020 FNP-FNS workshop. Several systems exploring different techniques have been introduced. Zheng *et al.* (2020) proposed a system that involved splitting the annual reports into their relative sections by parsing the Table of Contents and then applying a BERT-based classifier to determine which section to include in the final summary. On the other hand, Azzi & Kang (2020) implemented a similar approach with extracting the Table of Contents (TOC) from each annual report, but they made use of a Convolutional Neural Network (CNN) binary classifier using Keras, that classified all titles as either narrative or not, based on their presence in the reference summaries provided. On the other hand, Singh (2020) uses a different approach, that is based on combining both extractive and abstractive summarization methods by exploring pointer networks to extract important narrative sentences from the report, and then T-5 is used to paraphrase extracted sentences into a concise yet informative sentence.

The lack of gold datasets of financial documents and their corresponding summaries is the major bottleneck for applying any NLP or ML techniques for summarizing them.

3. Data

In this work, we focus on annual reports produced by UK firms listed on The London Stock Exchange (LSE). The dataset is built for the 2020 FNS Shared task (El-Haj *et al.*, 2020) and was made publicly available. The dataset has been extracted from UK annual reports published in PDF file format. UK annual reports are lengthy documents with around 80 pages on average, some annual reports could span over more than 250 pages, while the summary length should not exceed 1000 words. The training set includes 3,000 annual reports, with 3-4 human-generated summaries as gold standard that are 1000 words in length, while the validation set contains 363 documents with their corresponding gold summaries. For the evaluation process the test set of 500 files was provided. As the reports are provided in PDF file format, extracting structure is a challenging task. El-Haj *et al.* (2018) used the UK annual report's table of contents to retrieve the textual content (narratives) for each section listed in the table of contents. This dataset covers a set of 3,863 annual reports. We use the same data splits as the ones used in the FNS shared task.

3.1. Data Preprocessing

Due to the nature of the texts being originally extracted from PDF files, the organization and structure of the resulting TXT files is extremely noisy. Most of the established PDF to text conversion products on the market (i.e., pdf2text) generate highly noisy unstructured texts containing abbreviations, non-standard words, false starts, missing punctuation, missing letter case information, and other text disfluencies. Since extractive summarization is solely based on the information within the texts, it is important to be able to identify and process individual sentences or sections. We defined different steps for data cleaning including table removal, as tables usually containing numerical information that is not important in our context. We also removed information such as e-mails, contact information etc. Then we eliminate empty lines, and make sure each sentence is placed on separate lines for easier extraction.

We use regular expressions and the pretrained sentence level tokenizer, PunktSentenceTokenizer, to tokenize the text at the sentence-level, and place them on separate lines.

4. Methodology

In this section, we describe the methodology we explored to build several models for annual reports summarization. We follow two main approaches: 1) Sentence-based extractive summarization 2) Section-based extractive summarization. Before presenting these approaches, we will start by introducing the NLP technique we followed to encode sentences in a computational format.

4.1. BERT for Text Encoding

Sentence Encoding as Embeddings is an upstream task required in our task. To create each sentence embedding, we made use of the Bidirectional Encoder Representations from Transformers (BERT) language model (Devlin *et al.*, 2019). As several pretrained models were made publicly available, We decided to use the following three models, due to their relevance to our specific task at hand. *Bert-uncased-large* - the most downloaded model overall, *distilbart-cnn-12-6* (Sanh *et al.,* 2019) - the most downloaded model for summarization tasks, and finally, *FinBert* - a model pre trained on financial news that was annotated by 16 people with backgrounds in finance and business (Malo *et al.*, 2013). It is important to note that BERT has a maximum input length of 512 tokens at any one given time, meaning that it is unable to compute large amounts of data at once, which is why in this task, we computed individual sentence embeddings, and then computed whole section or document embeddings by taking an average of the relevant sentence embeddings.

4.2 Sentence-based Summarization

As discussed previously, an extractive summarization task typically consists of extracting the most relevant individual sentences from the original document. As a result, we decided to initially approach this task by creating sentence-based summaries, which included extracting individual sentences from the original documents, and only including the most relevant ones in the final summaries.

4.2.1 Sentence-based summarization as a Classification task

As an initial method, we formulated the extractive summarization task as a binary sentence classification task that assigns 1 to a given sentence if it is to be kept in the summary, and 0 if it is to be discarded. To do this, we needed to build a labelled training set. This process involved going through every sentence in all the annual reports in the training dataset and checking if it exists in the corresponding reference summary. If it exists, it is labelled as 1, otherwise it is labelled a 0. Then, we use the annotated training data to fine-tune the BERT model for 10 epochs, with a learning rate of 5e-5, a batch size of 32, and a max sequence length of 512. We pick the best fine-tuning checkpoint on the dev set and we create a sentence level classification model. We ran our model on the test set to classify the test sentences.

One major issue with this approach seemed to be the large class imbalance in the training dataset as annual reports usually contain thousands of lines. Out of these sentences, only 50 to 100 of them would be found in the corresponding reference summaries and labelled as 1, while the majority would be labelled as 0. This means that once the sentences were classified and fed into the BERT encoder, every sentence in the test set would be classified as 0, since the

training data had a major imbalance in the sentences classified as 0 vs. classified as 1.

4.2.2 CENTROID-SENTENCE-BASED Summarization

We also explored a centroid-based approach where a 'centroid' vector embedding representation of the whole document was initially determined by finding the mean of all the sentence embeddings. Then, to define the most important sentences in a document, we compute the cosine similarity between each individual sentence's embedding and the centroid vector embedding. Initially, the top 30 sentences with the highest similarity scores were kept in the final summary and the rest were discarded. This is because the reference summaries were 1000 words on average, which is approximately 30 sentences. However, it was apparent that the top sentences extracted were not coherent since they were not consecutive within the original document, so the resulting summaries were difficult to comprehend. Therefore, we decided to extract the top three sentences, and the seven sentences surrounding each of the top three sentences, which also made for a total of 30 sentences.

Although the resulting summaries seemed to be decent, due to the nature of the task being explored and the dataset used, once the summaries were evaluated, we realized that the reference summaries in the dataset were mainly extracted from whole sections within the original annual reports, rather than individual sentences. This meant that we had to redefine our extractive summarization task by switching from sentence-based summarization to section-based summarization, for us to cater to the nature of the dataset.

4.3 Section-based Summarization Approaches

4.3.1 Section Extraction

As defined by Litvak *et al.* (2020), there are typically 13 predefined narrative section titles found in an annual report[1]. We followed that same approach for section extraction.

4.3.2 SECTION-COSINE

We apply a similar approach like the one described in Section 4.2.2 for sentence-based summarization, which compares embeddings of individual sentences with the average embedding of the whole document. However, in this case, rather than comparing the centroid document embedding with sentence embeddings, we determined average vector embeddings for each section by averaging the embeddings of the sentences within a section.

Once the cosine similarity score for each section was determined, the first 1,000 words of the section with the highest similarity score were extracted from the original document and included in the final summary.

4.3.3 SECTION-CLUSTERING

To investigate relevant approaches within the domain, we explored the idea of clustering in extractive summarization. Clustering is typically performed on sentence-based summaries. Once the sentences are clustered, each cluster of sentence embeddings can be interpreted as a group of semantically identical sentences carrying the same information and whose meaning can be represented by only one sentence from the cluster (Gupta, 2020). Then the top sentence from each cluster is chosen by extracting the sentence with the top similarity score when compared to the centroid within its corresponding cluster.

However, rather than clustering individual sentences, whole section embeddings were clustered. The number of clusters in this case was 13 because we used a list of 13 predefined section titles that are usually found within the average annual report. A total of 14,808 sections from the training and validation datasets were clustered. Then, the most common section within each cluster was determined by looking at the frequency of their appearance in the reference summaries of the training and validation datasets, and the average embedding for each document in the testing dataset was compared to each of the 13 centroids (i.e.sections) using cosine similarity, to determine which of the centroids it is the most similar to. The top three closest centroids/sections for each document were then determined, and the first 1000 words of the top existing section in the original document were included in the final summary.

4.3.4 WEIGHTED-SECTION-CLUSTERING

After looking closely at the training set, we realized that some sections were selected more often than others. So, to further build on the clustering approach described above, rather than just taking the top available section as a summary candidate for a given document, we considered the frequency of each of the 13 sections in the training data and we define a weight for each section title. We labelled each reference summary with the name of the section that it was the most similar to.

This was done by comparing the reference section's average document BERT embedding to the extracted sections from the original document and used the same section name as the name of the section that it was closest to, in terms of cosine similarity. A frequency counter is then created to keep track of all the reference summaries that correspond to a certain section title. In doing so, we determined the weight of each section in the reference summaries. Once the top three sections for each document were determined through the clustering process, the existing section with the highest frequency weight was kept, and the top 1000 words of that section were included in the final summary.

[1] Section titles: ['"chairmans statement"', '"chief executive officer ceo review"', '"chief executive officer ceo report"', '"governance statement"', '"remuneration report"', '"business review"', '"financial review"', '"operating review"', '"highlights"', '"auditors report"', '"risk management"', '"chairmans governance introduction"', '"Corporate Social Responsibility CSR disclosures"']

	Rouge-1			Rouge-2			Rouge-L			Rouge-Sum4		
	R	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**
Baseline Methods												
TextRank	0.41	0.12	0.17	0.23	0.04	0.07	0.24	0.20	0.21	0.30	0.05	0.08
LexRank	0.34	0.27	0.26	0.19	0.11	0.12	0.21	0.26	0.22	0.25	0.12	0.14
Sentence-Based Experiment												
CENTROID-Sentence-based	0.37	0.21	0.26	0.14	0.06	0.08	0.30	0.19	0.23	0.23	0.08	0.11
FNS-2020 Shared Task Top Submissions												
SRIB2020-3	0.61	0.39	0.47	0.45	0.22	0.29	0.61	0.38	0.46	0.51	0.21	0.29
FORTIA-1	0.43	0.43	0.41	0.3	0.28	0.27	0.4	0.4	0.38	0.34	0.33	0.32
Section-Based Experiments (using 'bert-uncased-large' model)												
SECTION-COSINE	0.47	0.27	0.33	0.25	0.12	0.15	0.41	0.26	0.31	0.33	0.13	0.17
SECTION-CLUSTERING	0.46	0.29	0.33	0.24	0.13	0.16	0.40	0.28	0.32	0.32	0.14	0.17
WEIGHTED-SECTION-CLUSTERING	0.48	0.37	**0.38**	0.30	0.18	**0.21**	0.45	0.35	**0.36**	0.36	0.19	**0.22**

Table 1: ROUGE Scores of the several summarization models in addition to baseline models

5. Evaluation and Results

We intrinsically evaluated the summaries produced by the methods described in section 4.

5.1. Intrinsic evaluation

To evaluate the quality of the generated summaries, we use ROUGE (Lin, 2004), the de facto automatic summarization evaluation metric that compares an automatically produced summary against a set of reference gold summaries. Following the FNS Shared task setup, we evaluate our models using ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-SU4. We also report the results of two basic baseline methods: TextRank[2], an unsupervised text summarization technique inspired by the PageRank algorithm used primarily for ranking web pages in online search results (Mihalcea & Tarau, 2004), and LexRank[3], a graph-based unsupervised technique that relies on sentence connectivity (Erkan & Radev, 2004).

The results given in Table 1 compare the sentence-based and section-based systems to the performance of the baseline methods, in addition to the FNS-2020 top submissions.

5.2. Alternative Results – Validation Set

Since there are multiple reference summaries corresponding to each annual report, we decided to experiment with filtering the reference summaries down to the reference summary with the highest score when compared to our generated summaries in the validation set. The main intuition behind this evaluation is to demonstrate the vast variety in some reference summaries that were extremely unstructured and would penalize our models when evaluated. The alternative ROUGE scores following this evaluation approach are given in Table 2.

[2] https://github.com/summanlp/textrank

[3] https://pypi.org/project/lexrank/

	Rouge-1			Rouge-2			Rouge-L			Rouge-Sum4		
	R	P	F1	R	P	F1	R	P	F1	R	P	F1
CENTROID-Sentence-Based	0.41	0.37	0.39	0.10	0.09	0.10	0.40	0.36	0.38	0.16	0.15	0.15
SECTION-COSINE	0.56	0.50	0.53	0.34	0.30	0.32	0.55	0.49	0.52	0.38	0.34	0.36
SECTION CLUSTERING	0.57	0.50	0.53	0.34	0.29	0.29	0.57	0.50	0.53	0.38	0.33	0.35
WEIGHTED-SECTION-CLUSTERING	0.64	0.60	**0.62**	0.49	0.44	**0.46**	0.63	0.60	**0.62**	0.51	0.47	**0.50**

Table 2: Alternative Results – ROUGE scores of top gold summaries

5.3. Discussion

There are multiple observations that could instantaneously be made when looking at the results reported in Table 1. When comparing the sentence-based and section-based models, it is apparent that the section-based attempts perform much better when evaluated using the FNS-2020 shared task dataset. In fact, there seems to be an increase of 0.13 in the Rouge-L F1 Score. This means that section-based methods are much more relevant when exploring extractive summarization using this dataset. Furthermore, when analyzing the scores of the section-based approaches, the best-performing system turned out to be the WEIGHTED-SECTION-CLUSTERING. When this system is compared to the two baseline methods, there is a major increase in both recall and precision, which ultimately led to an increase in the Rouge-L F1-score, from 0.21-0.22 to 0.36. Referencing the top 2 shared task submissions, and specifically comparing the ROUGE-L F1-Scores, we can see that our proposed system, WEIGHTED-SECTION-CLUSTERING, would rank third amongst the top shared-task systems, with a score of 0.36. These shared task submissions had access to much more advanced section extraction tools, which could have positively impacted scores, which leaves room for improvement in that aspect of this research. Finally, when comparing the results to our alternative results, we see an immediate improvement in scores since the systems aren't being penalized for some of the reference summaries that do not match the generated system summaries. Rouge-L F1-Score increases from 0.36 to a much higher score of 0.61.

Our experiments clearly show that the quality of the dataset we use for this task is not perfect. The choice of the best sections included in the gold summaries was never intuitive and clear. There is no inter-annotator agreement reported. Every annotator would choose a section over another one in a very subjective manner.

6. Conclusion and Future work

In this paper, we explored the task of automatic extractive summarization of UK annual reports. We presented different models that fall under two main approaches: sentence-based summarization, which involves extracting the most informative sentences from the annual reports, and section-based summarization, which involves extracting the most informative section of the annual report. Due to the nature of the FNS-2020 dataset being used, section-based approaches performed better in terms of ROUGE. Our WEIGHTED-SECTION-CLUSTERING model yielded the best results when evaluated against the testing set, achieving a ROUGE-L F1-Score of 36% when evaluated against 3 reference summaries and 62% when only the closest reference summary is considered.

In the future, many limitations could be addressed through improving the accuracy of section identification and extraction. We also plan to tackle the task of sentence-based summarization in a more adequate setup by exploring the use of a dataset where summaries are built by extracting sentences in context, instead of full sections. Another important point is to conduct an extrinsic evaluation by including several Finance experts to evaluate the quality of the summaries generated automatically.

Acknowledgments

We would like to thank the FNS 2020 shared task organizers Mahmoud El-Haj and Ahmed Abu Raed for providing us with the dataset and helping us with carrying out the evaluation on the blind test set that is not available publicly.

References

[Abujar *et al*., 2017] Abujar, S., Hasan, M., Shahin, M. S. I., & Hossain, S. A. (2017). A heuristic approach of text summarization for Bengali documentation. *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*.

[Azzi *et al*., 2019] Azzi, A. A., Bouamor, H., & Ferradans, S. (2019). The FinSBD-2019 Shared Task: Sentence Boundary Detection in PDF Noisy Text in the Financial Domain. *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, 74–80.

[Azzi & Kang, 2020] Azzi, A. A., & Kang, J. (2020). FNS-Summarisation 2020 shared task: system description paper Extractive Summarization System for Annual Reports. *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*

[Berger & Mittal, 2000] Berger, A., & Mittal, V. O. (2000). Query-relevant summarization using FAQs. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics - ACL '00*.

[Chung & Yang, 2000] Chuang, W. T., & Yang, J. (2000). Extracting sentence segments for text summarization. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '00*.

[Devlin *et al.*, 2019] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*.

[El-Haj *et al*., 2018] El-Haj, M., Dr., Rayson, P., & Moore, A. (2018). Towards a Multilingual Financial Narrative Processing System. *The First Financial Narrative Processing Workshop: Proceedings of the 11th Edition of the Language Resources and Evaluation Conference*, 52–58.

[El-Haj *et al*., 2019] El-Haj, M., Rayson, P., Young, S., Bouamor, H., & Ferradans, S. (2019). Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019). *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019)*.

[El-Haj *et al*., 2020] El-Haj, M., Litvak, M., Pittaras, N., & Giannakopoulos, G. (2020, December). The Financial Narrative Summarisation Shared Task (FNS 2020). *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation* (FNP 2020).

[Erkan & Radev, 2004] Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, *22*, 457–479.

[EY, 2020] EY. (2020). Global Fintech Adoption Index 2019. Retrieved from EY website: https://fintechauscensus.ey.com

[Filippova *et al.*, 2009] Filippova, K., Surdeanu, M., Ciaramita, M., & Zaragoza, H. (2009). Company-oriented extractive summarization of financial news. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics EACL '09*.

[Fowler, 2018] Fowler, D. S. (2018). How many websites are there in the world? | tek eye. Retrieved April 17, 2021, from https://tekeye.uk/computing/how-many-websites-are-there

[Ghazali & Annum, 2010] Ghazali, Mohd., & Anum, N. (2010). The importance and usefulness of corporate annual reports in malaysia. *Gadjah Mada International Journal of Business*, *12*(1), 31.

[Gupta, 2020] Gupta, A. (2020, September 29). Understanding Text Summarization using K-means Clustering. Retrieved February 18, 2021. https://medium.com/@akankshagupta371/understanding-text-summarization-using-k-means-clustering-6487d5d37255

[Gupta & Legal, 2010] Gupta, V., & Lehal, G. S. (2010). A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, *2*(3).

[Horev, 2018] Horev, R. (2018, November 17). BERT Explained: State of the art language model for NLP. *Towards Data Science*. Retrieved from https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

[J.P. Morgan, 2020] J.P. Morgan. (2020b). Innovation with machine learning.

[Khant & Singh Mehta, 2018] Khant, A., & Singh Mehta, M. (2018). Analysis of Financial News Using Natural Language Processing and Artificial Intelligence. *Conference: International Conference on Business Innovation 2018*.

[Kunder, 2018] Kunder, M. de. (2018). WorldWideWebSize.com | The size of the World Wide Web (The Internet). Retrieved from https://www.worldwidewebsize.com

[Leidner, 2019] Leidner, J. L. (2019). Summarization in the Financial and Regulatory Domain. In *Trends and Applications of Text Summarization Techniques* (pp. 187–215).

[Lin, 2004] Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Conference: In Proceedings of the Workshop on Text Summarization Branches Out*.

[Lin & Och, 2004] Lin, C.-Y., & Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*.

[Litvak *et al.*, 2020] Litvak, M., Vanetik, N., & Puchinsky, T. (2020). Hierarchical summarization of financial reports with RUNNER. *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, 212–224. Barcelona, Spain.

[Malo *et al*., 2013] Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2013). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, *65*(4), 782–796.

[Mihalcea, 2004] Mihalcea, R. (2004). Graph-based ranking algorithms for sentence extraction, applied to text summarization. *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions.*

[Mihalcea & Tarau, 2004] Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Text. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404–411. Barcelona, Spain: Association for Computational Linguistics.

[Sanh *et al.*, 2019] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*.

[Singh, 2020] Singh, A. (2020). PoinT-5: Pointer Network and T-5 based Financial Narrative Summarisation. *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, 104–110.

[Stanton & Stanton, 2002] Stanton, P., & Stanton, J. (2002). Corporate annual reports: Research perspectives used. *Accounting, Auditing & Accountability Journal*, *15*(4), 478–500.

[”Training and fine-tuning”, n.d.] Training and fine-tuning. (n.d.). Retrieved April 25, 2021, from HuggingFace website: https://huggingface.co/transformers/training.html

[Xu *et al.*, 2013] Xu, W., Grishman, R., Meyers, A., & Ritter, A. (2013). A Preliminary Study of Tweet Summarization using Information Extraction. *Proceedings of the Workshop on Language in Social Media (LASM 2013)*, 20–29.

[Yang *et al*., 1999] Yang, J., Parekh, R., & Honavar, V. (1999). DistAl: An inter-pattern distance-based constructive learning algorithm. *Intelligent Data Analysis*, *3*(1), 55–73.

[Zheng et al., 2020] Zheng, S., Lu, A., & Cardie, C. (2020). SUMSUM@FNS-2020 Shared Task. *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, 147–151.