Big Data & AI Strategies

State of the Art NLP Tools with Practical Financial Applications such as Thematics, Summarisation, Q&A

We have compiled practical examples showing how you can use State of the Art (SotA) Natural Language Processing (NLP) and Artificial Intelligence (AI) tools to help with portfolio management tasks. The examples cover: Thematic Investing with Sentiment, Summarisation, Q&A, & vector Embedding. We also offer some insights into building efficient NLP pipelines for text processing, clustering, charting and de-duplication of text like repeated news topic from alternate sources

- **NLP for Thematic Investing**. Our prior research covers *Thematic* based Investment and Sentiment identification, which we have called SmartBuzz. We explore some concepts around thematic sentiment identification, such as Aspect Based Sentiment Analysis (ABSA) and Target Aspect Sentiment (TAS).
- **NLP Summarisation (TLDR).** Information overload can be addressed with modern NLP tools designed to summarise text – in as few as 3 lines of code (shown below). Modern transformer models make AI summarization highly effective. Only as few as 5 examples are required to help fine-tune a model to your editorial style. We present a case-study of REITs transaction summaries.
- **NLP Q&A.** Numerous models are now 'super-human' at answering questions according to the SQuAD 2.0 Leaderboard. We explore some Q&A models and highlight some of our concerns with some annual report and ESG examples.
- **NLP Embedding for Concept Search & Similarity.** NLP models can be used to compresses a document to high dimension vector space. This can be used for approximate concept or thematic search as well as visualistaion of clusters. Identification of duplicated news article with slightly different headline text from different vendors is made practical with vector-based cosine similarity.
- **NLP Pipelines.** We explore modern and efficient NLP pipelines made possible with HuggingFace tools as well as other processes to access large data via datasets with built-in distributed and parallel processing capabilities.
- **NLP Model Compression.** We presented NLP model compression, pruning and numerical precision down sampling techniques for faster inference.
- **DeepFin.** A tutorial will be hosted via zoom at on 20/Jan/2022. Please RSVP.

Summarise any text in just 3 lines of python code

```
from transformers import pipeline
summarizer = pipeline("summarization", model=None)  # Specify a model name, path, or None for default
print(summarizer("Some Long Text Here", min_length=10, max_length=20))
```

Source: HuggingFace, J.P. Morgan.

BigData, AI & NLP Strategies

Berowne Hlavaty AC
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com
Bloomberg JPMA HLAVATY <GO>
J.P. Morgan Securities Australia Limited

Robert Smith, PhD
(61-2) 9003-8808
robert.z.smith@jpmorgan.com
J.P. Morgan Securities Australia Limited

Evan Hu
(852) 2800-8508
evan.hu@jpmorgan.com
J.P. Morgan Securities (Asia Pacific) Limited

Peng Cheng, CFA
(1-212) 622-5036
peng.cheng@jpmorgan.com
J.P. Morgan Securities LLC

Ayub Hanif, PhD
(44-20) 7742-5620
ayub.hanif@jpmorgan.com
J.P. Morgan Securities plc

Arun Jain
(1-212) 622-9454
arun.p.jain@jpmorgan.com
J.P. Morgan Securities LLC

Chief Global Market Strategist

Marko Kolanovic, PhD
(1-212) 622-3677
marko.kolanovic@jpmorgan.com
J.P. Morgan Securities LLC

See page 40 for analyst certification and important disclosures, including non-US analyst disclosures.

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

NLP for Thematic Investing

We have written extensively on thematic investing using our SmartBuzz framework. For example, see SmartBuzz notes on Earnings, News, and Sentiment. There are essentially two parts to most thematic investment tasks: (1) identifying the theme, and (2) assigning sentiment. We describe both of these tasks in detail.

Basic Sentiment Analysis

We note that recent models such as the transformers package from HuggingFace makes sentiment analysis possible with just 3 lines of code, as shown below.

Figure 1: Sample Code to Classify Sentiment

```
from transformers import pipeline
classifier = pipeline("text-classification")
classifier("The advances in NLP over the past decade are amazing")
>>>>  [{'label': 'POSITIVE', 'score': 0.9998743534088135}]
```

Source: HuggingFace, J.P. Morgan.

However, a problem with standard sentiment analysis is that dense writing often packs more than one topic into a sentence as shown in the example below.

Figure 2: Example of dense sentence with multiple topics mentioned:

"In the Corporate & Investment Bank, Global IB fees were up 52% driven by a surge in M&A activity and our strong performance in IPOs, Markets revenue was very strong overall and down just 5% compared to a third quarter record last year, as Continued normalization in Fixed Income was partially offset by a strong performance in Equities."

Source: J.P. Morgan 3Q21 announcement

n-gram

In the fields of computational linguistics and probability, an n-gram is a contiguous sequence of n items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. The n-grams typically are collected from a text or speech corpus. An n-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram"; size 3 is a "trigram". Larger sizes are sometimes referred to by the value of n, e.g., "four-gram", "five-gram", and so on.

Source: www.definitions.net

In such a situation a sentiment score of "NEUTRAL" (or "MIXED" if using our Sentiment training datasets) will be returned, but clearly ***different aspects of the sentence have different sentiment***.

One approach we have explored is to identify the target word, then measure the sentiment of the surrounding n-gram words. While this technique is relatively computationally cheap, it ignores more complex sentence structure.

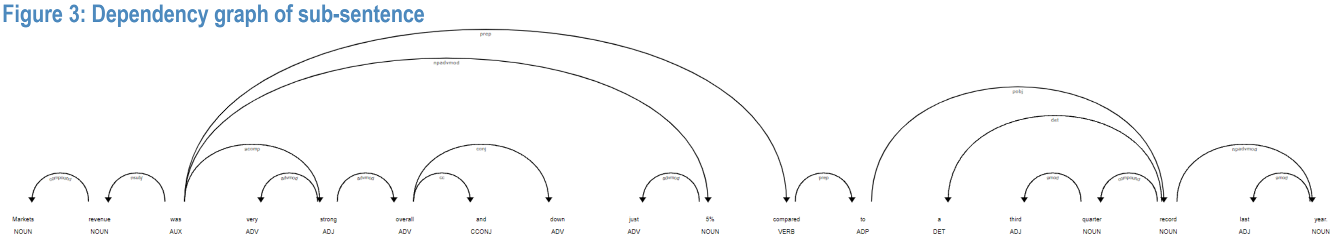
Another approach is to split sentences by punctuation, which could work for simple language sentences but could miss some negation and other complex structures.

A more complex approach would be to split the text into smaller independent chunks, sometimes called spans. These could include just a single sentence or perhaps use a dependency graph to parse the sentence and focus on the part of the graph that relates to the aspect of concern. Finer (shorter) spans could be created by splitting at commas as well as periods. As a result, each [sub-] sentence could be processed

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

independently. Note that longer spans have richer context information, so a model will have more information to consider.

Figure 3: Dependency graph of sub-sentence

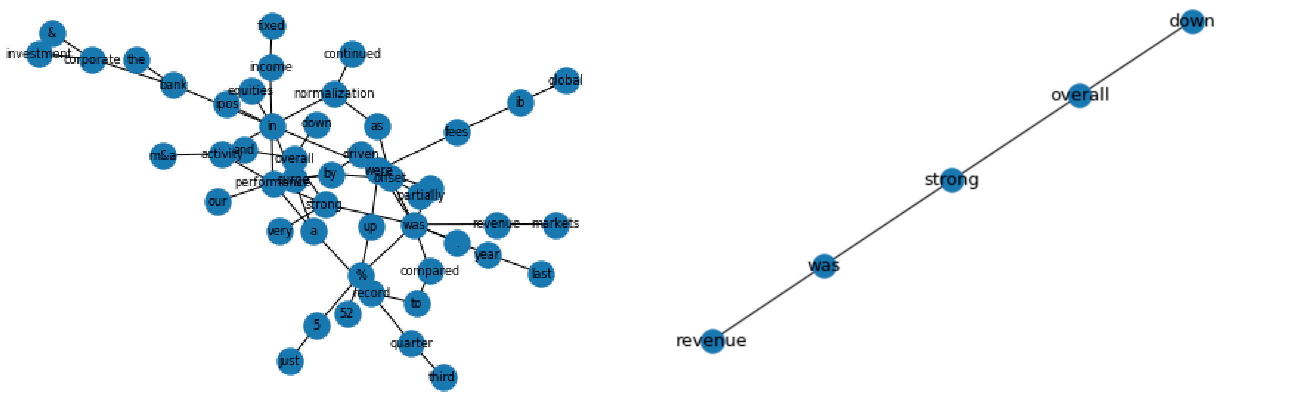


Source: J.P. Morgan. Spacy

Shortest Dependency Path (SDP)

We can also use network graphs to visualize the text… but the repeated use of words creates multiple nodes which can become confusing.

Figure 4: Network Graph of words in sentence and SDP



Source: J.P. Morgan.

Drilling in to the "revenue" sub-graph we have the shortest path between revenue and down_5% as:

'revenue', 'was', 'strong', 'overall', 'down' …[5%]

However a problem with this method is that one needs to specify the two ends of interest, so while it might be easy to identify theme words as the starting word, deciding on the second token is not so easy, limiting the usefulness of this approach.

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

Sentence Segmentation for Sentiment Specificity

A potential and useful approach becomes a relatively simple split of complex sentences by comma. Note: we had to add one additional comma (highlighted in the code snippet below as ,) to separate Fixed Income from Equities. Without the additional segmentation the Fixed Income from Equities sentence segment is analysed as a single block and tagged as MIX or NEU depending on the model.

Figure 5: Sample Code to Segment Sentences

```
doc = """In the Corporate & Investment Bank, Global IB fees were up 52% driven by a surge in M&A activity
         and our strong performance in IPOs, Markets revenue was very strong overall and down just 5%
         compared to a third quarter record last year, as continued deterioration in Fixed Income was,
         partially offset by a strong performance in Equities."""
doc = [ d for d in doc.replace(",",".").split(".") if d !=''   ]

JPM_Senti = pipeline("text-classification", model='./MPNet_Model')
sentiment = JPM_Senti(doc)

for s, d in zip(sentiment, doc):
    print(f"{s['label']} {d}")

>>> OUTPUT:

NEUT In the Corporate & Investment Bank
POS  Global IB fees were up 52% driven by a surge in M&A activity and our strong performance in IPOs
POS  Markets revenue was very strong overall and down just 5% compared to a third quarter record last year
NEG  as continued deterioration in Fixed Income was
POS  partially offset by a strong performance in Equities
```

Source: J.P. Morgan QDS.

An alternative to using commas as a sentence segment tool is to use stop words, however this is likely to result in more and shorter sentence segments. A more robust approach might be to generate a list of common topic separators, like '… offset by…' and '… compared to …' or '… contrast…' and use these to segment sentences as well as commas.

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

Aspect Based Sentiment Analysis

Thematic investing can be aided by what is known as Aspect-Based Sentiment Analysis in academia. The technique offers a more nuanced approach to sentiment scoring than document and sentence scores. This becomes especially important when trying to measure the sentiment of a particular aspect/theme/company in a mixed sentence.

In the example explored earlier (see above, Figure 2), the overall sentiment is mixed, but *positive* for Equities [Revenue] and *negative* for Fixed Income. We explore this paragraph again but this time with ABSA. We have specified a target aspect "Equities" and "Fixed Income" and the model estimated the sentiment for this particular topic and highlights the most/least significant words that contribute to the sentiment scores.

The ABSA model seems to focus on the 'deterioration' wording around Fixed Income and gives a moderate probability of a negative sentiment for Equities, despite highlighting 'strong performance in equities' at the end of the sentence.

Figure 6: Example of Incorrect Negative Sentiment for Equities Revenue

```
Sentiment.negative for "Equities"
Scores (neutral/negative/positive): [0.066 0.617 0.317]
```

Importance 1.00 in the corporate & investment bank , global ib fees were up 52 % driven by a surge in m & a activity and our strong performance in ipos , markets revenue was very strong overall and down just 5 % compared to a third quarter record last year , as continued deterioration in fixed income was partially offset by a strong performance in equities .

Source: J.P. Morgan QDS.

The ABSA model correctly focused on the 'deterioration' wording around Fixed Income and gives a high probability of a negative sentiment as a result, highlighting 'continued deterioration in fixed income' as the key part of the sentence.

Figure 7: Example of Correct Negative Sentiment for Fixed Income Revenue

```
Sentiment.negative for "Fixed Income"
Scores (neutral/negative/positive): [0.008 0.979 0.014]
```

Importance 1.00 in the corporate & investment bank , global ib fees were up 52 % driven by a surge in m & a activity and our strong performance in ipos , markets revenue was very strong overall and down just 5 % compared to a third quarter record last year , as continued deterioration in fixed income was partially offset by a strong performance in equities .

Source: J.P. Morgan QDS.

This example highlights that the ABSA model can be useful in simple sentences, but it is easily confused by compound sentences. We simplified the example and can achieve a positive sentiment for Equities as shown below.

Figure 8: Example of Correct Positive Sentiment for Equities Revenue – with modified input sentence

```
In the Corporate & Investment Bank, Global IB fees were up 52% driven by a surge in M&A activity and our strong performance in
IPOs, and strong performance in Equities.
Sentiment.positive for "Equities"
Scores (neutral/negative/positive): [0.044 0.349 0.607]
```

Importance 1.00 in the corporate & investment bank , global ib fees were up 52 % driven by a surge in m & a activity and our strong performance in ipos , and strong performance in equities .

Importance 0.69 in the corporate & investment bank , global ib fees were up 52 % driven by a surge in m & a activity and our strong performance in ipos , and strong performance in equities .

Source: J.P. Morgan QDS.

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

Dividend Aspect Example

ABSA can successfully assign the correct positive sentiment to both the company and the concept of dividends in another sample… however beware that each of the aspects is analysed separately, increasing processing time.

Figure 9: ABSA Analysis for "Dividends" and "BCE" Aspects

```
[3]: recognizer = absa.aux_models.BasicPatternRecognizer()
     nlp = absa.load(pattern_recognizer=recognizer)
```

```
In [19]: text = ("Binh Duong Construction and Civil Engineering Joint Stock Company (HOSE:BCE), " \
             "Vietnam's 31st largest Engineering & construction company by market cap, " \
             "is currently the 19th highest dividend yielding stock in the VN Index of 278 stocks. ")
         completed_task = nlp(text=text, aspects=['BCE', 'dividend'])
         stock, dividend = completed_task.examples
```

```
In [23]: absa.summary(stock)
         absa.display(stock.review)
```

```
Sentiment.positive for "BCE"
Scores (neutral/negative/positive): [0.054 0.004 0.942]
```

Importance 1.00 binh duong construction and civil engineering joint stock company (hose : bce) , vietnam ' s 31st largest engineering & construction company by market cap , is currently the 19th highest dividend yielding stock in the vn index of 278 stocks .
Importance 0.66 binh duong construction and civil engineering joint stock company (hose : bce) , vietnam ' s 31st largest engineering & construction company by market cap , is currently the 19th highest dividend yielding stock in the vn index of 278 stocks .
Importance 0.59 binh duong construction and civil engineering joint stock company (hose : bce) , vietnam ' s 31st largest engineering & construction company by market cap , is currently the 19th highest dividend yielding stock in the vn index of 278 stocks .
Importance 0.57 binh duong construction and civil engineering joint stock company (hose : bce) , vietnam ' s 31st largest engineering & construction company by market cap , is currently the 19th highest dividend yielding stock in the vn index of 278 stocks .
Importance 0.41 binh duong construction and civil engineering joint stock company (hose : bce) , vietnam ' s 31st largest engineering & construction company by market cap , is currently the 19th highest dividend yielding stock in the vn index of 278 stocks .

```
In [24]: absa.summary(dividend)
         absa.display(dividend.review)
```

```
Sentiment.positive for "dividend"
Scores (neutral/negative/positive): [0.073 0.02  0.906]
```

Importance 1.00 binh duong construction and civil engineering joint stock company (hose : bce) , vietnam ' s 31st largest engineering & construction company by market cap , is currently the 19th highest dividend yielding stock in the vn index of 278 stocks .
Importance 0.38 binh duong construction and civil engineering joint stock company (hose : bce) , vietnam ' s 31st largest engineering & construction company by market cap , is currently the 19th highest dividend yielding stock in the vn index of 278 stocks .
Importance 0.22 binh duong construction and civil engineering joint stock company (hose : bce) , vietnam ' s 31st largest engineering & construction company by market cap , is currently the 19th highest dividend yielding stock in the vn index of 278 stocks .
Importance 0.18 binh duong construction and civil engineering joint stock company (hose : bce) , vietnam ' s 31st largest engineering & construction company by market cap , is currently the 19th highest dividend yielding stock in the vn index of 278 stocks .
Importance 0.18 binh duong construction and civil engineering joint stock company (hose : bce) , vietnam ' s 31st largest engineering & construction company by market cap , is currently the 19th highest dividend yielding stock in the vn index of 278 stocks .

Source: J.P. Morgan QDS. LexusNexus

MPNet Sentiment and ELI5 for Diagnostics

This particular example highlights some of the complexity faced with NLP Sentiment Scores for compound sentences. Here the ABSA model is returning a **NEG** prediction for "Equities" but the comma split sentence segment returned a **POS** using MPNet on "… *Markets revenue was very strong overall and down just 5% compared to a third quarter record last year*…"

Digging into this example we note that for the "Revenue" aspect the inclusion of words: 'markets' (-3.267), 'revenue' (-2.104), 'was' (-5.695) have negative score contributions. The keywords contributing to these scores are highlighted below.

Note that ELI5 does not reverse colours for negative scores, so for the REVENUE diagnosis in output 2 below with a negative score of -2.0 (and low probability of 0.1) we see that 'revenue was … down' in green are the biggest contributors to the negative score while 'down… compared' in red detract from this negative score.

Figure 10: ELI5 Analysis for MPNet Sentiment Prediction of "Revenue" Aspect

```
***** Running Prediction *****
  Num examples = 5000
  Batch size = 64
```

y=Markets (probability **0.035**, score **-3.267**) top features

Contribution?	Feature
-0.820	<BIAS>
-2.447	Highlighted in text (sum)

markets revenue was very strong overall and down just 5% compared to a third quarter record last year

y=revenue (probability **0.113**, score **-2.014**) top features

Contribution?	Feature
-0.826	<BIAS>
-1.188	Highlighted in text (sum)

markets revenue was very strong overall and down just 5% compared to a third quarter record last year

y=was (probability **0.003**, score **-5.695**) top features

Contribution?	Feature
-0.479	<BIAS>
-5.216	Highlighted in text (sum)

markets revenue was very strong overall and down just 5% compared to a third quarter record last year

y=very (probability **0.848**, score **1.999**) top features

Contribution?	Feature
+1.539	Highlighted in text (sum)
+0.461	<BIAS>

markets revenue was very strong overall and down just 5% compared to a third quarter record last year

Source: J.P. Morgan QDS.

While the last segment for 'very' has the highest probability (0.85) for a positive score (of near 2) … flagging 'strong' and 'just' as positive contributors but negative for 'down', which was partially negated by 'just', as shown in the last section of the figure above.

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

Coreferences

Co-References occur frequently in text. Often a person or entity is introduced in full at the start of a paragraph, then referenced by a short form (he, she, it) later. Using a coreference tool such as neuralcoref it is possible to identify all coreferences and optionally replace them in the text. Coreference systems are typically designed to solve the Winograd problem, nicely explained in this medium.com article.

Figure 11: Example News Article Text:

"While the process is still in the early days, Dimon said the bank has already begun to reap some rewards by putting additional resources into nabbing new business in international markets. He cited Asia as being a region where the bank had one of its best years ever in 2020."

Source: Business Insider

Figure 12: Sample Code to identify Coreferences

```
import spacy
import neuralcoref
nlp = spacy.load('en_core_web_md')
neuralcoref.add_to_pipe(nlp)
doc = nlp(u"While the process is still in the early days, Dimon said
the bank has already begun to reap some rewards by putting additional
resources into nabbing new business in international markets. He cited
Asia as being a region where the bank had one of its best years ever in
2020.")
print( doc._.coref_clusters)
```

Source: J.P. Morgan.

The resulting coreferences have been highlighted in the original text above, and can be seen as a list of "coref clusters" output from the spacy+neuralcoref model:

- Dimon: Dimon, He
- the bank: the bank, the bank, its

Unfortunately, coreferences have been designed for proper nouns, and do not work so well for themes and other nouns like Revenue or Interest. It is possible that additional model training/finetuning might resolve this, but this task is left to the enthusiastic reader.

There is also a problem with coreference models, related to a lack of quality training data and low out-of-sample performance on other data. This problem has not been fully addressed by a single source, but several researchers have touched on the issue, specifically in regard to the limitations of OntoNotes (a.k.a. CONLL 2012 dataset) as a benchmark and the inadequacies of coreference systems that are trained with OntoNotes as their sole benchmark when they are applied to other datasets.

Juntao Yu's work ("A Cluster Ranking Model for Full Anaphora Resolution") in particular has focused more than others on the issues with OntoNotes and the coreference systems based on it. "Moving on from OntoNotes" by Xia and Van Durme is also worth reviewing for recommendations on alternative datasets.

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

Target Aspect Sentiment (TAS)

The next advance in this space looks at 3x components, using Named Entity Recognition (NER) to identify a *Target*, processing an *Aspect* and reporting the *Sentiment*, as discussed in "Target-Aspect-Sentiment Joint Detection for Aspect-Based Sentiment Analysis" and on GitHub. However, this package is not on PyPi and is beyond the scope of this report.

Aspect-Controllable Opinion Summarization (ACOS)

A recent paper (Nov 2021) on Aspect-Controllable Opinion Summarization offers an interesting alternative worth exploring further. The model creates targeted summaries per aspect (thematic) which could then be scored with a sentiment engine.

Figure 13: Aspect-Controllable Opinion Summarization Abstract

Recent work on opinion summarization produces general summaries based on a set of input reviews and the popularity of opinions expressed in them. In this paper, we propose an approach that allows the generation of customized summaries based on aspect queries (e.g., describing the location and room of a hotel). Using a review corpus, we create a synthetic training dataset of (review, summary) pairs enriched with aspect controllers which are induced by a multi-instance learning model that predicts the aspects of a document at different levels of granularity. We fine-tune a pretrained model using our synthetic dataset and generate aspect-specific summaries by modifying the aspect controllers. Experiments on two benchmarks show that our model outperforms the previous state of the art and generates personalized summaries by controlling the number of aspects discussed in them.

General
The room was clean and comfortable. The staff was very friendly and helpful. It was a great location, just a short walk to the beach. There wasn't much to do in the area, but the food was good.
Location
The location was great, right on the Boardwalk, and close to the Venice beach.
Rooms
The room was very clean and the bathroom was very nice. The bathroom had a large separate shower. There was a TV in the room.
Location and **Rooms**
The location is great, right on Boardwalk, and the beach is very nice. The room was very clean and the bathroom was very nice and the shower was great.
Cleanliness, **Location**, **Room**, and **Service**
The staff was very friendly and helpful. The room was very clean, and the bathroom was very nice. It was a great location, right on the beach.

Table 1: General and aspect-specific summaries generated by our model for a hotel from the SPACE dataset. Aspects and aspect-specific sentences are color-coded.

Source: DOI: 10.18653/v1/2021.emnlp-main.528

Conclusions on Thematic Investing

The current SotA Sentiment for whole sentences and longer is largely solved, but the detailed aspect level or thematic sentiment problem is still a work in progress where proprietary models have the potential to outperform. We have worked extensively on this problem with our SmartBuzz research. We note there are two recent open-source projects that are worth further investigation in this area: TAS and ACOS. For now, we recommend scoring sentence segments with a traditional sentiment engine.

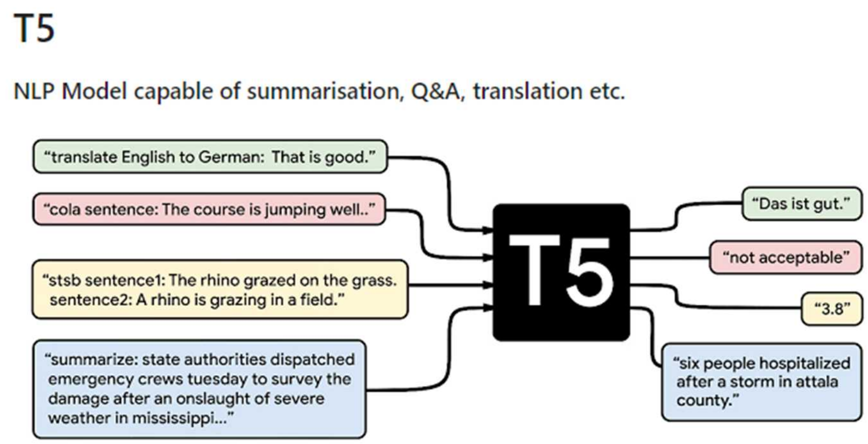
Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

NLP Summarisation

Information overload is a common problem across many different areas within our daily lives, and especially so when it comes to financial information. To help reduce the task we can turn to text summarization techniques. More information on a range of alternatives can be found at MachineLearningPlus.

Last year we were excited by the possibility of text summarization when we saw the documentation for the T5 model.

Figure 14: Some of the T5 Models capabilities



Source: HuggingFace, Microsoft

Unfortunately, the T5 base model responses were underwhelming when we tried.

Figure 15: Some summaries can be underwhelming or miss the point

```
".com..com..com..com..com..com..com..com..com..com..co
m..com..com..com..com..com..com..com..com..com..com..c
om..com..com..com"

"I am pleased to report that the year ended in a very
challenging environment for our industry........"
```

Source: Source: J.P. Morgan, T5 v1.1 base.

Thankfully these have been improved (greatly!) with the help of some more recent NLP models, and can be further refined with additional fine-tuning. On HuggingFace we note there are currently 76 summarization models, some focused on: news, medical publications, patents, meetings, title generation, legal and financial (1) domains.

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

Text Summarisation Examples

Producing reasonable summaries using NLP tools has come a long way, both in terms of model performance, but also in terms of ease of use. Below we explore a few examples – again in as few as 3 lines of code.

Figure 16: Sample Code to Summarise Text

```
from transformers import pipeline
summarizer = pipeline("summarization", model='human-centered-summarization/financial-summarization-pegasus')
print(summarizer("Some Long Text Here",...))
```

Source: HuggingFace, J.P. Morgan.

How to Judge Summary Text

Part of the problem for machine created summaries, is how do you judge two competing sentences that summarise the same page of text? Even if the exact same words are used, but in a different order the summary might not be valid, so customized ranking algorithms are needed to score summary engines. One example "Choice of Plausible Alternatives" or ROUGE (doesn't handle synonyms) or METEOR (designed for translation). Most DNN models use some form of Label Smoothing to help reduce model overconfidence as discussed at NeurIPS 2019. NLP models typically use the probability of each token which is smoothed and compared to the true label tokens to calculate the error.

However, we have found that some manual verification of the quality of text summarization is *absolutely* required. Also, the length of summary is only *somewhat* controllable.

For example, we have a few long articles in Appendix1. We have used three different models to create the summaries as shown. As you can see, the base T5 model is not ideal and gave us pause. Recently the newer models released to the HuggingFace model repository and show much improved summarisation. These examples were run *before* any additional fine-tuning by us (models were downloaded as is).

Figure 17: Comparison of Three AI Models for Text Summarisation

T5: t5-base	DistillBart: sshleifer/distilbart-cnn-12-6 (default HuggingFace model)	Pegasus: human-centered-summarization/financial-summarization-pegasus
prosecutors say the marriages were part of an immigration scam. if convicted, barrientos faces two criminal counts of "offering a false instrument"	Liana Barrientos pleaded not guilty to two criminal counts of "offering a false instrument for filing in the first degree" In total, she has been married 10 times, with nine of her marriages occurring between 1999 and 2002.	Prosecutors say Barrientos was part of an immigration scam. She is believed to still be married to four men at one time, prosecutors say
on June 25, Alibaba DAMO Academy (the R&D branch of Alibaba) announced they had built M6. it is a large multimodal, multitasking language model with 1 trillion parameters. the model was intended for multimodality and multitasking.	Alibaba DAMO Academy (the R&D branch of Alibaba) announced they had built M6, a large multimodal, multitasking language model with 1 trillion parameters. The model was intended for multimodality and multitasking, going a step further than previous models towards general intelligence.	Alibaba's M6 is 50 times larger than GPT-3's size. These achievements will have far-reaching positive consequences for the AI community
the objective of this paper is to conduct a performance comparison of five deep learning models. the attention and DCN perform best with wavelet or FFT signal, whereas some other models perform better with no data preprocessing.	Wind power experienced a substantial growth over the past decade especially because it has been seen as one of the best ways towards meeting climate change and emissions targets by many countries. The accuracy of wind forecasts is a key element for the electric system operators, as it strongly affects the decision-making processes.	Novel and cutting-edge deep learning models are investigated. Performance comparison of FFT, DCN, DFF, RNN, LSTM models
a year ago, our industry was dealing with the shockwaves arising from the pandemic. a year ago, our industry was dealing with the confluence of demand contraction. a year ago, a price war between OPEC and Russia resulted in global oversupply of oil.	Underlying profit of $447 million was smaller than usual but was still a very sound result, given the scale of the challenges. The reported loss, of just over $4 billion, reflects the major writedowns of our assets announced in July.	Underlying profit smaller than usual but still very sound: CEO. Loss reflects major writedowns of assets announced in July: chief executive I would like to take this opportunity to thank you for your support over the past 12 months

Source: J.P. Morgan QDS

Fine-Tuning Text Summarisation to Match Your Style

After examining some of the base models, we decided to test the impact of 'fine-tuning'. Given the financial focus of the Pegasus model we used this model for our customization experiments.

J.P. Morgan REITs team

Richard Jones
Radhika Wadhwa
Solomon Zhang
Annabelle Atkins
Blake Im (Intern)

We had 10x news articles with handwritten summaries covering recent Real Estate transactions with thanks to our REITs team. The REITs team have a need to review all transactions within the past month so they can summarise the market activity. Their writing style has a particular focus on deal size and yield levels for the rental income.

We used the first 5 for training and the results show the pick-up of yield focus in the summaries. This took about 1 hour (500sec for 5 rows) to train. Summarising an article takes about 1.5s each. In most cases the fine-tuned model performed better, picking up the yield or cap-rate as an important part of the summary and adopting the editorial style quite well. The fourth example was the only example where the raw model did better than the fine-tuned model (which introduced repetition).

However, caution is required with the 3rd example citing 5% instead of 4%! We suspect that fine-tuning with different formatted numbers introduced this problem.

Figure 18: Comparison of Human, Transformer and Fine-Tuned AI Models for Text Summarisation

Human Summary	AI Model (without Fine-Tuning)	AI (with Fine-Tuning)
Fawkner Property has acquired Stockland Cairns from SGP for $146m, representing book value and a 6.75% yield. The shopping centre spans ~50,000sqm and is anchored by Coles, Woolworths, Big W, Target and BBC Cinemas. {Red text is additional Information not in article}	Fund manager will use the cash to help bankroll a new Essential Services fund. Stockland has been shedding non-core assets as part of a plan to slash costs and focus on high-quality,	Stockland has sold its Cairns asset, on 11.13 hectares at Earlville, for $146m, representing a 6.75 per cent yield. The transaction comes two months after Fawkner acquired Mackay's Mount Pleasant Centre from Vicinity Centres for $162.5m.
SCP has acquired Delacombe Town Centre in Ballarat from Troon Group for $112m, representing an implied fully let yield of 5.3%. The sub-regional asset was completed in 2017, spans 19,098sqm and is anchored by a Woolworths, Kmart and Dan Murphy's.	Delacombe Town Centre is a strong performing Woolworths, Kmart and Dan Murphy's anchor. The deal for the 19,098 square metre sub-regional mall on Glenelg Highway was struck on an implied fully let yield of 5.3 per cent.	SCA Property Group has acquired Delacombe Town Centre from Troon Group for $112m, with an implied fully let yield of 5.3 per cent. The 19,098 square metre sub-regional mall is anchored by Woolworths and Kmart-anchored, with Dan Murphy's and a small number of specialty stores.
Cook Property Group acquired the Bunnings Hervey Bay development from Wesfarmers for $59m, representing an initial yield of 4.0%. The 17,421sqm store will be sold with a 10 year lease to Bunnings, with options until 2080.	The deal is a new benchmark for strong leased retail assets, with a 4 per cent yield. Cook Property Group bought the store from Bunnings in a deal brokered by Stonebridge, Savills and	New Zealand investor Cook Property Group has acquired the Hervey Bay Bunnings Warehouse development from DXS for $58.6m (5% cap rate), with options extending until 2080. The transaction was the second highest yield in Australia this year, after the Eastgarden Bunnings in Sydney, which was acquired by Newmark for $75m.
Woolworths Balwyn was sold to an undisclosed purchaser for $46m.	A major supermarket investment in Melbourne's Balwyn has transacted for $45.7 million, around 20% over expectations. The property was marketed by Vinci Carbone and included a 10-year lease to provide security for the store	Vinci Carbone's Frank Vinci and Joseph Carbone marketed the 383 Whitehorse Road Woolworths property at 383 Whitehorse Road, Melbourne from 1 July 2018. The property spans 18,000sqm of 280,000 of 280,000 of 280,000 of 280,000 of out-of-town retail space.
Coles Hawthorn was sold to a private investor for $25m, representing an extremely tight initial yield of 1.9%.	A new-format Coles supermarket in Melbourne's inner suburbs sold for $24.5 million on a record low yield of 1.9 per cent. The sale of the Hawthorn store smashed previous low return benchmark, set only last month, when a Woolworths supermarket on Canterbury Road in nearby Blackburn South sold for $29.3 million.	New-format Coles Local at 689-699 Glenferrie Road in Hawthorn has been sold to a local investor for $24.5m, with a yield of just 1.9 per cent. The sale smashed previous low return benchmark set only last month, when a Woolworths supermarket on Canterbury Road in nearby Blackburn South sold for $29.3m.

Source: J.P. Morgan QDS

We expect further examples will help fine-tune the model further, but for now suggest that *these models will help produce quality draft notes that still require human verification*.

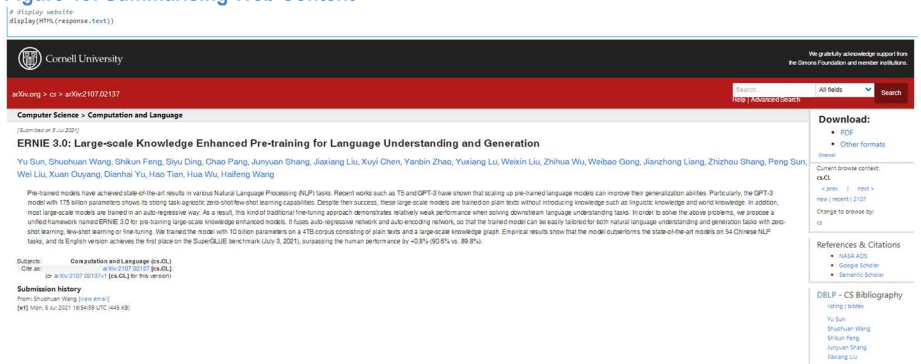
Summarisation of Raw Online Content

For our BigData and AI Summaries reports we need to review and summarise hundreds of articles about the latest industry and academic innovations.

Simple Content

Sites like ArXiv are a great source of reports on innovation that offer a relatively simple summarization challenge. Essentially our models are able to identify the abstract as containing the most relevant text to summarise.

Figure 19: Summarising Web Content



Source: J.P. Morgan QDS. ArXiv.org

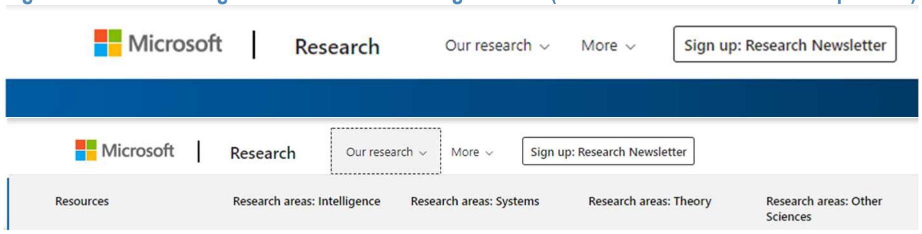
{'summary_text': 'Pre-trained language models have achieved state-of-the-art results, but still lack fine-tuning. ERNIE 3.0 unifies auto-regressive, auto-encoding, zero-shot learning, and fine-tuning approaches to natural language understanding'}

Problem Content

Some sites have considerable data in the headers, such as Microsoft's Research Blog.

The base summarisation model finds text in the hidden drop-down options (e.g. "Our research:" as shown below). Resulting in {'summary_text': 'U.S.-based research is carried out in the following areas: intelligence, human-computer interaction, Quantum computing, data platforms, analytics, and more. Research areas: Systems Research, Systems Data platforms, analytics, human-computer interaction'}

Figure 20: Summarising Web Content – with long headers (sometimes hidden behind dropdowns)



Source: J.P. Morgan QDS. Microsoft.com

In this case we need to skip down 400 words to find the start of the article. The risk here is that some models such as Pegasus have a 512-token limit (approx. 400 words) and so we could miss the relevant text.

Note a token can represent a whole word, punctuation, or a number if they are included in the dictionary. Rare words are broken up into multiple tokens generally representing syllables. For this reason 400 words can take about 500 tokens.

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

For these articles we need to use the DIV processing to skip down to the article or body text. However, every website uses different div names (some randomly generated strings) so we scan for a few common div names and hope to find the body text using this technique. Alternatively, we can list each DIV and select the DIV with the most text.

```
text = ""
divs =['#body', '#articleBody','#articleText', '#main', '#main-content']
for bodyDiv in divs:
    print(bodyDiv)
    if soup.select(bodyDiv):
        text = soup.select(bodyDiv)[0].text
    if text: break
```

Once the appropriate text is obtained, summarization is trivial. Example based on Microsoft's and nVidia's Multimodel Transformers article.

```
{'summary_text': 'We propose an approach to reduce the
size of multimodal transformers up to 97 percent. This
allows us to train our model end-to-end on 30 seconds of
video'}
```

We caution that it is obviously a difficult challenge to know in advance which websites offer simple content and which will require additional parsing to identify the core text of the article!

Fine Tuning Summary Pipeline

We have taken one of the best performing AI models for text summarization (Pegasus) and fine-tuned it using over 200 human generated summaries. Training on a 4-GPU AWS server took about 6 hours (<USD$20). Summaries can then be created on our hold-out samples in about 1 second each (GPU) or 4 seconds (CPU). Note the Pegasus model is over 2Gb.

Comparison of Pre/Post Fine-Tuning

Below we compare the raw Pegasus model against one that we have fine-tuned with 160 training samples. Training was run for 10 epochs before we used the model to summarise some out-of-sample news reports. Training samples included summaries of single articles and concepts.

As you can see from the examples below, care needs to be taken in relying upon these models, but the performance and quality of output has improved dramatically over the past 6 months.

Figure 21: Comparison of Human vs A.I. generated Summaries

Human Summary	AI Summary (Before Fine-Tuning)	AI Summary (After 10 epoch Fine-Tuning with 160 samples)
Nvidia also unveiled new software solutions such as Megatron, a PyTorch-based library for large-scale transformer models, and Jarvis, a framework for building conversational agents.	Megatron is a PyTorch-based framework for training giant language models. Riva has also announced new achievements for Riva, a fully accelerated conversational AI {Mentioned Riva instead of nVidia as source for announcement}	Megatron is a PyTorch-based framework for training giant language models based on the transformer architecture. Larger language models are helping produce superhuman-like responses and are being used in applications such as email phrase completion, document summarization and live sports commentary. {Excellent, but focused on main product skipping Jarvis and Riva products mentioned later in article}
NVIDIA update the desktop with NVIDIA RTX A4000, claimed to be the most powerful single-slot GPU.	U.S.-based company has been at the forefront of artificial intelligence for more than a decade. Company has been at the forefront of the field for more than a decade {unhelpful use of 'Company' and no actual information about the product introduced}	The A4000 is the most powerful single-slot GPU for professionals, delivering real-time ray tracing, AI-accelerated compute, and high-performance graphics to your desktop. {Excellent}
Xwing, a California startup, is test-flying an autonomous pilot system aboard cargo aircraft with an eye toward crewless commercial flights in 2022. A suite of models reads sensor data while the plane is in motion. When the models detect another plane or an obstacle, they funnel the information to a rules-based flight control system, which adjusts course	*Planes can fly itself, start with a pilot in the cockpit. ProBankruptcyCentral BankingPrivate EquityStrategic IntelligenceVenture CapitalCMO TodayLogistics Report* *{Poor information content and web tags used as text}*	*To Build a Plane That Can Fly Itself, Start With a Pilot in the Cockpit - ~~WSJ Skip to Main ContentSkip to Search Dow Jones, a News Corp company focused on creating and distributing authoritative and engaging content and other products and services.~~* *{News Article Headline used – and WSJ website text included – need to filter input}*
IBM Research published a paper detailing the design of an AI chip optimized for low precision training and inference. This new AI core and chip can be used for many new cloud to edge applications across multiple industries.	World's first energy efficient AI chip built with 7nm technology. Self-driving cars, text to speech, delivery drones all in the works {Generic information content}	The world's first energy efficient AI chip built with 7nm technology. This chip technology can be scaled and used for many commercial applications from large-scale model training in the cloud to security and privacy efforts by bringing training closer to the edge and data closer to the source. {Good summary but would have liked to see "low precision training" included – a personal preference?}
The cross-platform framework offers the most compelling web development experience. Flutter websites are only interpretable by people, not by search engines, so that's going to have a huge impact on how people search for and find information on your website.	Flutter is a cross-platform framework for building websites. It's been around for a few years now, but could be about to take off {Summary misses the point and product name}	We've had Flash, we've had Silverlight, all competing technologies that have attempted to shave off a slice of the browser market, to let the developer use a different technology to create a website. But none of them have taken off. {Summary misses the point and product name}
Qlib (0.7.0) is an AI-oriented quantitative investment platform, which aims to realize the potential, empower the researcher and extract the value of AI technologies in quantitative investment.	Qlib is an AI-oriented quantitative investment platform, which aims to realize the potential, empower the research, and create the value of AI. {Good. Summary missed 'er' from researcher – uncommon usage}	Qlib is an AI-oriented quantitative investment platform, which aims to realize the potential, empower the research, and create the value of AI technologies in quantitative investment. It contains the full ML pipeline of data processing, model training, back-testing; and covers the entire chain of quantitative investment: alpha seeking, risk modeling, portfolio optimization, and order execution. {Good, but summary still missing 'er' }
Shapash, an open-source python library that helps Data Scientists to make their Machine Learning models more transparent and understandable by all. It makes it easier to share and discuss the model interpretability with non-data specialists: business analysts, managers, end-users	shapash makes Machine Learning models transparent and understandable by everyone. {Succinct}	Shapash is a Python library which aims to make machine learning interpretable and understandable by everyone. It provides several types of [output] that display explicit labels that everyone can understand. {Good summary, missing "output" word}
Combining Azure cloud with AI & ML Tools, Microsoft is making health data accessible, easy to use for healthcare workers and organizations as they need it.	Health Cloud & Platform features FHIR data management with a single, consistent platform. Health data in the cloud with *FHIR data management with a single, consistent platform* {Repeated output lacking insight}	Microsoft Next's Health Cloud & Platform team's vision is to make health data accessible, easy to use and available to all healthcare workers and organizations as they need it. Data interoperability is at the core of our service and we work to enable data workloads that can light up analytics, AI and ML once we help bring your data into Azure in a way designed for security and compliance. {Good but long summary, missing product name}

Source: J.P. Morgan QDS. HuggingFace, Pegasus

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

Sample Code for Text Summarisation of Web Content

Remember to check the sites robot.txt file before web-scraping.

```
import requests
from IPython.core.display import HTML
from bs4 import BeautifulSoup
from transformers import pipeline

# Setup summarization pipeline
model = 'human-centered-summarization/financial-summarization-pegasus'
summarizer = pipeline("summarization", model = model,)

def url_summary(summarizer, url, min_length=30):
    response = requests.get(url, proxies=None)
    # optional: display website
    # display(HTML(response.text))
    # extract text content
    soup = BeautifulSoup(response.text, 'html.parser')
    return summarizer(soup.text, truncation=True, max_length=130, min_length=min_length)[0]

url='https://arxiv.org/abs/2107.02137'
url_summary(summarizer, url, min_length=50)

>>>  {'summary_text': 'Pre-trained language models have achieved state-of-the-art results,
but still lack fine-tuning. ERNIE 3.0 unifies auto-regressive, auto-encoding, zero-shot
learning, and fine-tuning approaches to natural language understanding'}
```

Conclusions on Text Summarization Models

As we have shown with various examples above, Text Summarization has recently been largely solved with the performance and quality of output having improved dramatically over the past 6 months. However, we caution that care needs to be taken in relying upon these models, especially with poorly formatted web-scraped source text. These models make for a useful draft that should be edited by a human before conclusions are drawn.

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

NLP Q&A

Information overload can also be addressed for specific problems, especially if the summary can be delivered in the form of a question. While we continue to focus this discussion on HuggingFace tools, with over 100 hosted models for English language Q&A and also 18 models specifically for tabular Q&A (all based on tapas).

However, we note many of the hosted models are based on popular tools like BERT and have been fine-tuned for squad 1 or 2 benchmarks by various contributors. Further we note that the pipeline method only supports 33 models with labels "….ForQuestionAnswering" suffix such as Google's BigBird Pegasus.

There are other alternatives like https://haystack.deepset.ai/ and RAG-DAG (FaceBook) that we explore further below and find use for as a vector encoding/embedding model.

Open Domain Q&A

Somewhat simpler problem as the answer is drawn from (and often highlighted in) the provided context.

Also, a lot of earlier NLP models didn't know how to politely *not* answer questions they couldn't answer. For this there is a part of SQuAD 2.0 that addresses unanswerable questions. NLP models have been 'super-human' since mid-2019, with models often based on ensembles.

For larger databases, only the query is supplied at runtime and a *retriever* fetches relevant documents (i.e. context) for a *reader* to extract answers from. There is a summary of these systems here: How to Build an Open-Domain Question Answering System.

Closed Book Q&A

Some modern and large language models like T5 or GPT-3 have memorised some facts during pre-training and can generate an answer without explicit context (the "closed-book" part is an analogy with humans taking exams, answering questions from memory). These models are mentioned in the above blog post, while this T5 paper is also informative: How Much Knowledge Can You Pack Into the Parameters of a Language Model?

Concern

Some models need careful monitoring as they can answer a question with words or numbers not included in the context.

Figure 22: Sample Code to Answer Questions

```
from transformers import pipeline
question_answerer = pipeline("question-answering")

context = """Extractive Question Answering is the task of extracting an answer from a text given a
question. An example of a question answering dataset is the SQuAD dataset, which is entirely based on
that task. If you would like to fine-tune a model on a SQuAD task, you may leverage the
examples/pytorch/question-answering/run_squad.py script."""

result = question_answerer(question="What is extractive question answering?", context=context)
print(f"Answer: '{result['answer']}', score: {round(result['score'], 4)}, start: {result['start']},
end: {result['end']}")
>>>> Answer: 'the task of extracting an answer from a text given a question', score: 0.6177, start: 34,
end: 95
```

Source: HuggingFace, J.P. Morgan QDS.

We caution that some of the results from generative models might include 'facts' that are not based on the context – making their use for financial models problematic.

Mixed Mode learning is also becoming prevalent, such as models that can perform Q&A with video input, like this one called Violet.

Currently we see most of the Q&A models have been trained on SQUAD v1 or v2 (which added unanswerable questions). The extractive answering models are typically posting short-form answers of one number or one to a few words.

These are good for specific questions like "What was the profit" or "When will you achieve net zero emissions?" and can be surprisingly good for vague questions like "What is the carbon reduction plan" (Answer: 'net zero for our direct emissions by 2050.'). However, for a detail minded analyst we might argue a more correct answer is "to decarbonize coal-fired power-plants with the use of hydrogen as ammonia and/or transition to LNG". To be fair to the AI Models, the text used is a bit vague on this topic though.

Examples are taken from a recent AGM for Woodside.

Figure 23: Sample Questions and Answers

Question	Answer	Score
What was the profit for 2021	'$447 million'	0.6553
When will you achieve Net Zero Emissions	'2050'	0.9301
What is your carbon reduction plan	'net zero for our direct emissions by 2050.'	0.4157

Source: J.P. Morgan QDS. Company Report

Figure 24: Sample Code to Answer Questions

```
result = question_answerer(question="What was the profit for 2021?", context=context)
print(f"Answer: '{result['answer']}', score: {round(result['score'], 4")
>>> Answer: '$447 million', score: 0.6553,
```

Source: HuggingFace, J.P. Morgan QDS.

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

Fact Based Data Extraction

Given the problems with (some) large language models potentially generating answers from a mix of extraction and (unattributed) memory, we feel that exploring some alternative methods might prove necessary.

Network Graph Extraction (for Facts and Relationships)

Another related method to perform Q&A is to extract all the facts into a network graph (or sql database) which can be used to answer questions with certainty.

An example on medium shows how to use coreferences to build a relationship graph.

Another paper looks at "Improving Multi-hop Question Answering over Knowledge Graphs using Knowledge Base Embeddings".

We also discussed a similar technique during one of our 4th DeepFin sessions with Microsoft "NLP and Knowledge Graphs" which is available on GitHub.

Tabular Data Extraction

Google have released TaPas (also on HuggingFace) and TABERT to help understand tabular data, with recent updates to enhance the models performance on GitHub.

Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) offers the promise of streamlining the creation of natural language processing models by combining the strengths of open-book models and the memory of closed-book models. In a RAG system, essentially the entire corpus is embedded in a dense vector space. Next a question is encoded in the same vector space, and nearest neighbors to the question are retrieved using a fast search (FAISS). Candidate text blocks are passed to a generator to create the output.

Multiple Choice

There might be certain situations where your problem has a multiple-choice style answer to choose between. For these problems, the QuaRTz dataset is available to train models with. An example is available in HuggingFace model repo here.

Q&A Case Study – ESG Answers for Net Zero goals

For Q&A problems the SQUAD training dataset is the most appropriate. SQUAD v2 is an enhanced training dataset with approximately one third of questions un-answerable from the provided text. We have searched over the HF model repo and found a few candidate models to test, most pre-trained on SQUAD 1 or 2.

Please note that Scope 1 represents direct company activities, Scope 2 captures Indirect *Upstream* effects such as purchase of electricity, while Scope 3 captures all other *Downstream* effects such as the use of products and *Upstream* effects like travel, transport and other purchases and investments etc. See PlanA.Earth for more details.

While simple Q&A problems are largely solved according to the SQUAD 2 leaderboard, we thought it would be more helpful to examine real-world questions about carbon emissions that are hard to obtain from current ESG databases.

Figure 25: ESG Questions, Guide Answers and Notes

Question	Guide Answer	Notes
When will the company achieve net zero emissions?	2050	2021 cited by some models as this is the first year mentioned in the document header on all pages.
What is the scope 1 direct emission reduction target?	15% reduction by 2025, 30% reduction by 2030, net zero by 2050	2 interim targets are mentioned along with net-zero goal of 2050
What year will scope 1 target be done by?	2025, 2030, 2050	2050 is the ultimate year but interim mentions worthy candidates
What is the scope 2 upstream emission reduction target?	15% reduction by 2025, 30% reduction by 2030, net zero by 2050	2 interim targets are mentioned along with net-zero goal
When will scope 2 emissions be reduced?	2025, 2030, 2050	2050 is the ultimate year but interim mentions worthy candidates
What is the scope 3 indirect emission reduction target?	na	Not mentioned
What investment has been made to reduce carbon emissions?	• planting some 3.6 million trees in Western Australia • opening of the Woodside Building for Technology and Design at Monash University, • investment in the HyNET consortium, which is building hydrogen fueling stations in Korea	Subtle answers may not be obvious match to the question
What is the current scope 1 direct emission levels?	na	Not mentioned

Source: J.P. Morgan QDS. Company Reports.

This particular report from Woodside Petroleum AGM from 2021. The PDF was converted to text and used as the 'Context' for our NLP Q&A Models.

We note that the problem is particularly challenging in this example because of different targets mentioned along the path to net-zero.

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

Figure 26: HuggingFace Question Answering Models Tested

Model Label	Model Size	Description	Timing/ Question	Score of 8	Results
Primer/bart-squad2	406Mn	BART model trained for SQUAD 2	32s	6	Very good results (See below)
bert-large-cased-whole-word-masking-finetuned-squad	333Mn	Traditional BERT model with whole word cased tokens	25s	5	First 5 questions correct. Q6 cited "innovation in the design of facilities" which is a near miss.
madlag/bert-large-uncased-wwm-squadv2-x2.63-f82.6-d16-hybrid-v1	284Mn	BERT large model trained for SQUAD 2	20s	3	Answered only 2050 for all questions except Q7 on investments.
ahotrod/electra_large_discriminator_squad2_512	334Mn	Electra large model trained on SQUAD2	22.2s	2	Poor results generally
kiri-ai/t5-base-qa-summary-emotion	223Mn	Base T5 model fine-tuned for SQUAD 2	14s	4/2	Better than T5-Base. Q1, Q2, Q4, Q5 partially correct. Q3 referenced 2021 AGM. Q6 ref Scope 1 targets not Scope 3. Q7 incorrectly referenced indigenous communities. Q7 A made up answer.
google/t5-base	223Mn	Early leader in Q&A and multi-mode learning. Needs more fine-tuning to be effective.	15s	3/2	Poor, answers miss parts of words or wrong years / goals mentioned.
ahotrod/albert_xxlargev1_squad2_512	206Mn	albert model trained for SQUAD2	158s	1	Poor but some answers in the right vicinity.
rahulchakwate/albert-xxlarge-finetuned-squad	206Mn	Albert model trained for SQUAD	176s	1	Q2 correct but other answers poor quality
google/bigbird-pegasus-large-arxiv	577Mn	Large Google model trained on academic articles	38s	0	Poor performance with 'long-term relationships' in most answers. Only supports short context.

Source: J.P. Morgan QDS. HuggingFace

Some models found the first year mentioned (2021 AGM) and answered 'When' questions with this, rather than the correct target year for net-zero. The other questions are tricky to answer. '15% reduction by 2025 and 30% reduction by 2030 and net zero by 2050'

None of the models predicted 'Not Answerable' for any question.

Below we tabulate the answers from the best performing model.

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

Figure 27: ESG Questions and Answers from Primer/bart-squad2 Model

Question	Guide Answer	Model Answer	Context
When will the company achieve net zero emissions?	2050	2050	... In 2020, we outlined our decarbonisation pathway as we aim for net zero by 2050 , targeting a 15% reduction in net equity Scope 1 and 2 emissions by 2025 a...
What is the scope 1 direct emission reduction target?	15% reduction by 2025, 30% reduction by 2030, net zero by 2050	15%	... ed our decarbonisation pathway as we aim for net zero by 2050, targeting a 15% reduction in net equity Scope 1 and 2 emissions by 2025 and a 30% reductio...
What year will scope 1 target be done by?	2025, 2030, 2050	2025	... y 2050, targeting a 15% reduction in net equity Scope 1 and 2 emissions by 2025 and a 30% reduction by 2030. We think our net emissions have peaked and wi...
What is the scope 2 upstream emission reduction target?	15% reduction by 2025, 30% reduction by 2030, net zero by 2050	30%	... geting a 15% reduction in net equity Scope 1 and 2 emissions by 2025 and a 30% reduction by 2030. We think our net emissions have peaked and will be decl...
When will scope 2 emissions be reduced?	2025, 2030, 2050	2050	... gy, taking steps along our pathway to net zero for our direct emissions by 2050 . When we met last year, we talked about how Woodside was managing the cha...
What is the scope 3 indirect emission reduction target?	na	net zero	... the last AGM, our major markets in Japan, China and Korea have all adopted net zero targets of their own. We're working with our customers in those countries ...
What investment has been made to reduce carbon emissions?	• planting some 3.6 million trees in Western Australia • opening of the Woodside Building for Technology and Design at Monash University, • investment in the HyNET consortium, which is building hydrogen fueling stations in Korea	HyNET consortium	... nizing coal-fired power generation in Japan. Through our investment in the HyNET consortium, which is building hydrogen fueling stations in Korea, we are gaining valua...
What is the current scope 1 direct emission levels?	na	net zero	... cantly, we matured our climate strategy, taking steps along our pathway to net zero for our direct emissions by 2050. When we met last year, we talked about ...

Source: J.P. Morgan QDS. Company Reports. Note Answer 5 cites year 2050, but the context relates to Scope 1 direct emissions rather than Scope 2 per the question.

Conclusions on Q&A Models

'Open Book Q&A' is largely a 'solved' problem and is arguably ready to use. However, we would caution that the SQUAD training answers are concise often single valued answers, so we recommend returning the answer and surrounding context for verification purposes.

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

NLP Embedding for Thematic Search

A recent paper from FacebookAI discusses how to perform entire document embedding. The original paper was aimed at tackling the knowledge-intensive NLP tasks (think tasks a human wouldn't be expected to solve without access to external knowledge sources), RAG models are seq2seq models with access to a retrieval mechanism providing relevant context documents at training and evaluation time.

For large Q&A problems, a RAG model encapsulates two core components: a question encoder and a generator. During a forward pass, the input is encoded with the question encoder and passed to the retriever to extract relevant context documents. The documents are then prepended to the input. Such contextualized inputs are passed to the generator. Read more about RAG at ArXiVs.

We have found the RAG document encoder to be a good method to embed documents in vector space for fast (yet approximate) concept or *Thematic Searches*. In this example documents are encoded as 768 long vectors.

FAISS is then used to quickly perform a hierarchical scan of nearest neighbors to a seed term embedding…. The assumption being that in embedding space, the closest matches to a search term will contain news articles about that topic.

We can scan the 16.6mn paragraphs quickly… two seconds to find the first 10,000 matches.

The advantage of the RAG process is that documents can be embedded during ingestion and searched for matching themes at any time. Searching can be performed at scale using numerical algo's and can also scale across large databases in parallel. The disadvantage is that the results are not exact text matches, so some documents might not be returned even if they contain a key word. An advantage, especially with English language, is that many words can have different meanings causing false-positive matches with text searches. Here vector embedding is better able to distinguish the different meanings of words.

Poor Example

When we used takeover related search terms such as 'takeover', 'target' as well as a longer M&A related term: "m&a deal announced target acquirer makes a bid offer for in shares and script merger", we found the nearest neighbors search correctly matches *less* "Target Inc" company reports but misses many M&A related documents that a plain text search matched.

There was also a large subset of false matches on many companies that 'increased stake in' a subsidiary.

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

Figure 28: M&A clustered documents identified by vector encoding model – flagging poor match quality.

M&A : m&a deal announced target acquirer makes a bid offer for in shares and script merger

title	text	matches	scores	cluster	sentiment	Comp_n	doc_id
16:31 EDT Box closes $500M investment led by KKR	Box (BOX) announced that it has closed the previously announced $500M investment led by KKR (KKR). In connection with the closing of the investment, John Park, Head of Americas Technology Private Equity at KKR, has been appointed to the Box Board, effective immediately. Box anticipates using substantially all of the proceeds to...	41	89.6636	M&A	NaN	2	LNM_44908581487
The Wall Street Journal: Private-equity giant TPG may go public through IPO or SPAC	TPG, with nearly $100 billion in assets under management, hasflirted with an IPO[1]multiple times, only to end up balking while rivals forged ahead.Blackstone GroupInc. \n BX, \n -1.83%[2], \n Apollo Global ManagementInc. \n APO, \n -0.34%[3], \n KKR & Co. \n KKR, \n -1.89%[4] \n and Carlyle Group Inc. \n CG, \n -0.73%[5] \n we...	45	89.3263	M&A	NaN	4	LNM_45284812953
Associates' Defaults Haunt NTPC's Jhabua Bid	with one from Adani Group and Gautam Thapar's Avantha Holdings.	10	86.5703	M&A	NaN	9	LNM_45208109766
Forge Closes Over $150M in Funding, Adds New Investors Amid Another Record-Breaking Quarter; Following FINRA approval of Forge and SharesPost to operate as a single broker dealer, Deutsche Börse, Temasek and Wells Fargo Strategic Capital invest in Forge's	information includes trades conducted through SPFC, Forge Markets, and Emerson Equity, LLC (Member FINRA/SIPC and a broker dealer for SharesPost, Inc.) in 2011.\n\nView source version on businesswire.com: https://www.businesswire.com/news/home/20210504005348/en/	18	86.4324	M&A	NaN	5	LNM_44843780767
comScore (NASDAQ:SCOR) Earns Buy Rating from Analysts at Craig Hallum	comScore (NASDAQ:SCOR) Earns Buy Rating from Analysts at Craig Hallum	7	85.3871	M&A	NaN	7	LNM_45265346751
Money transfer firm Wise confirms direct listing plans for early July	and 'B' shares in order to "support Wise's focus on its mission as it transitions into the public markets". \n\nClass 'B' shares hold 9 votes per share, are strictly non-transferable and, amongst other voting right cancellation events, expire on the fifth anniversary of any listing. The 'B' shares will not be listed. The compan...	57	84.0924	M&A	NaN	7	LNM_45235301716
Juniper Networks Chairman Scott Kriens' value of investment increases by $9.1 million in the past quarter Tuesday June 29, 2021	past\n\nThe present value of $1000 invested a year ago is$1,335\n\n PV$1000 1-week 1-month 1-year\n JNPR.NYSE 1,001 1,058 1,335\n S&P 500 Index 1,016 1,014 1,426\n\n7.5 Price Change % (1 Mo, 3 Mo, 1 Yr)\n\n1-Year price change for Juniper Networks was 29.1%. Compared with the S&P 500 Index which rose 42.6% in the year, the rela...	28	84.078	M&A	NaN	6	LNM_45272582483
Aaron Wealth Advisors Increases Stock Holdings in General Mills, Inc. (NYSE:GIS)	and grain snacks. \n\nSee Also: What is the CAC 40 Index? \n\nWant to see what other hedge funds are holding GIS?Visit HoldingsChannel.com to get the latest 13F filings and insider trades for General Mills, Inc. (NYSE:GIS). \n\nInstitutional Ownership by Quarter for General Mills (NYSE:GIS)	24	83.4805	M&A	NaN	7	LNM_44865461343
Edge Capital Group LLC Reduces Stock Position in Alphabet Inc. (NASDAQ:GOOG)	Momentum Indicator: Relative Strength Index	2	83.0936	M&A	NaN	7	LNM_45240668337
Former ICE CFO Looks Back on Dozens of Deals Made During His Tenure; Scott Hill retired last week after more than 14 years as finance chief of the marketplace operator	revenue is now over $7 billion in revenue.\n\nWSJ: CFOs at public companies these days manage more than just the numbers.\n\nMr. Hill: Any...financial person can make the numbers work and make the model tell you the output you want. But you would never walk into the boardroom with a deal if you don't understand the strategy tha...	48	80.3328	M&A	NaN	8	DNA_WSJO000020210520eh5k0025t
Week's Best: Merrill Lynch Hangs Up on Cold Calling - Barrons.com	one.\n\nSibling advisors go indie. A brother-and-sister team that advised on more than $1 billion in assets at J.P. Morgan has launched a registered investment advisor firm with help from Dynasty Financial Partners. Called the Invictus Collective, the new RIA has offices in Milwaukee, Chicago, and Miami.\n\nThe complexities of ...	67	79.7305	M&A	NaN	5	DNA_DJDN000020210528eh5s002up
Reaches Gender Parity on Its Board of Directors as Shareholders Approve Election of Directors, Appointment of Auditors and Say on Pay	utilities have either a female CEO or Board Chair.\n\nAppointment of Auditors\n\nShareholders of the Corporation approved the appointment of Deloitte LLP as the Corporation's auditors to hold office until the close of the next Annual Meeting of Shareholders. //st\n\n# Votes For % Votes For # Votes Withheld % Votes Withheld\n\n2...	45	79.5703	M&A	NaN	2	LNM_44862300773

Source: J.P. Morgan QDS.

Positive Example

Search term: "carbon neutral carbon low emission emissions net zero co2 decarbonize green house gas dioxide neutrality" Correctly ignores "Carbon Black" pigment and company name, as well as "Carbon Fiber" materials and "Carbonated" beverages. Along with the carbon mentions we note the positive matches included the related topics: energy, oil, fuel, and LNG.

Figure 29: Carbon clustered documents identified by vector encoding model

title	Kinder Morgan, Inc.. Energy Transition Fireside Chat 10 Key Takeaways; NatGas Infra Suited to Complement Renewables. KMI US. GPS-3563683-0	Air Liquide. Takeaways from the Sustainability Day. AI FP. GPS-3687875-0	Kinder Morgan, Inc.. Energy Transition Fireside Chat 10 Key Takeaways; NatGas Infra Suited to Complement Renewables. KMI US. GPS-3563683-0	BASF SE. CMD takeaways and model update. BAS GR. GPS-3692414-0	Diamondback Energy. Inside the Corner Suite: 10 Key Takeaways from Fireside Chat with CEO Travis Stice and CFO Kaes Van't Hof. FANG US. GPS-3715051-0
text	output, especially through linepack management...	when it expects absolute CO2 emissions to star...	1 emissions only, or direct emissions under co...	The focus of BASF's CMD (last Fri) was to disc...	a commitment to Scope 1 carbon emission neutra...
GPS_ID	GPS-3563683-0	GPS-3687875-0	GPS-3563683-0	GPS-3692414-0	GPS-3715051-0
scores	112.881	84.8362	101.888	110.488	113.52
feiss	552900	585623	552901	587898	611042
matches	20	14	13	13	13
cluster	Carbon	Carbon	Carbon	Carbon	Carbon
net	0	0	0	1	0
co2	0	0	0	0	0
carbon	0	2	0	2	3
emission	9	5	4	4	4
green	0	0	0	0	0
neutrality	0	1	0	0	1
emissions	9	3	3	4	3
gas	1	0	4	0	0
low	1	1	2	0	1
zero	0	1	0	1	0
dioxide	0	0	0	0	0
house	0	0	0	0	0
neutral	0	1	0	1	1
decarbonize	0	0	0	0	0

Source: J.P. Morgan QDS.

Figure 30: Carbon clustered documents identified by vector encoding model – flagging carbon mentions as well as energy, oil, fuel, LNG.

title	text	matches	scores	cluster	sentiment
Kinder Morgan, Inc.. Energy Transition Fireside Chat 10 Key Takeaways; NatGas Infra Suited to Complement Renewables. KMI US. GPS-3563683-0	output, especially through linepack management, while offering lower emissions, make it indispensable to meet the county's energy demand for some time. \n\n\n (2) A not so convenient truth: renewables scope 3 emissions often ignored, window dressing actual emissions footprint. When considering total emissions, George highlights the often mischaracterized and overlooked concept of scope 3 emissions, indirect emissions associated with the manufacturing of purchased products or other lifecycle emissions (per NREL). When factoring in scope 3 emissions associated with constructing a renewable facility, natural gas becomes competitive with renewables on the basis of cumulative emissions. Often, companies will benchmark against Scope	20	112.881256	Carbon	NEG
Air Liquide. Takeaways from the Sustainability Day. AI FP. GPS-3687875-0	when it expects absolute CO2 emissions to start reducing. It targets a 33% reduction in its Scope 1 + 2 emissions by 2035 vs. 2020 and carbon neutrality (i.e. zero CO2 emissions) by 2050. The focus on absolute CO2 emission cuts is welcome given the relatively high CO2 intensity and aggressive absolute CO2 emission cuts also announced by other companies in the sector.\n\n\nHydrogen revenue target of ~€6bn pa by 2035 vs. ~€2bn today: Assuming this happens and the company's usual target 10%+ post-tax ROCE on €8bn investment, the incremental "low-carbon" hydrogen-related EBIT in 2035 could be ~€1bn pa (JPMe) with	14	84.836243	Carbon	POS
Kinder Morgan, Inc.. Energy Transition Fireside Chat 10 Key Takeaways; NatGas Infra Suited to Complement Renewables. KMI US. GPS-3563683-0	1 emissions only, or direct emissions under control of the operating entity (i.e., electric cars). However, as George notes, natural gas combined cycle total emissions remain similar, if not lower, when factoring in battery construction and replacement. The research supports the often-overlooked low CO2 emission from burning natural gas.\n\nFigure 2: Natural Gas Emissions Compared to Renewable Alternatives – Slide 19\n\n\n\nSource: KM analysis, EIA, Lazard, NREL, and North American Electric Reliability Corporation (NERC).\n\n(3) Demand tailwinds seen for natural gas. KMI continues to see natural gas demand growing from a number of factors, approximating a 1-2% CAGR consistent with PPI through 2040.	13	101.887756	Carbon	POS
BASF SE. CMD takeaways and model update. BAS GR. GPS-3692414-0	The focus of BASF's CMD (last Fri) was to discuss the roadmap of the company to achieve net zero Scope 1 + Scope 2 CO2 emissions by 2050. The company also accelerated its 2030 target to achieve 25% reduction in CO2 emissions vs. 2018-base compared to the previous target of carbon-neutral growth (i.e. flat CO2 emissions) till 2030. This aligns BASF with some other European chemical companies which have also announced targets of absolute CO2 emissions reduction over this decade – thereby aiming to reduce their carbon footprint which we see as a positive to improve the ESG profile of	13	110.487991	Carbon	POS
Diamondback Energy. Inside the Corner Suite: 10 Key Takeaways from Fireside Chat with CEO Travis Stice and CFO Kaes Van't Hof. FANG US. GPS-3715051-0	a commitment to Scope 1 carbon emission neutrality, and the company plans to become the "low-cost carbon operator." FANG mentioned that it would purchase carbon offset credits to offset remaining emissions and that it intends to eventually invest in income-generating projects that will more directly offset remaining Scope 1 emissions (investing with a third party through one of FANG's subsidiaries was specifically mentioned). FANG hopes to reduce methane intensity by >70% and Scope 1 GHG intensity by >50% from 2019 levels by 2024, with continued reductions in flaring (down ~90% from 2019), which impacts >50% of 2019 Scope 1 emissions.	13	113.520126	Carbon	POS
BP. Feedback from CFO meeting: Upstream CMD in Oman to focus on portfolio runway and strong CFFO momentum; reiterate OW. BP/ LN. GPS-2842584-0	the company accepts that the shift will take some time. Dr. Gilvary also discussed the company's Reduce, Improve, Create (RIC) framework with targets including: i) zero net growth in operational emissions out to 2025; ii) 3.5 Mte of sustainable GHG emissions reductions by 2025; iii) 0.2% methane intensity and holding it below 0.3%; iv) improvements through lower emissions gas, more efficient and lower carbon fuels, lubricants. Additionally, with $200-500m allocated in 2019 for venture spending on businesses relating to energy transition, the company was keen to highlight the importance to generate a return from these investments, and that they were	13	114.182144	Carbon	NEUT
Air Liquide. Takeaways from the Sustainability Day. AI FP. GPS-3687875-0	Air Liquide held its Sustainability Day virtual meeting yesterday. The key messages: 1) acceleration of the de-carbonization efforts to move from carbon intensity reduction (but still rising absolute CO2 emissions) until 2025 to absolute CO2 emissions reduction from 2025 onwards to reach carbon neutrality by 2050; 2) as was largely expected, the company plans to accelerate its investment for "low-carbon (green + blue hydrogen)" hydrogen roll-out with a target to triple its hydrogen related revenue to ~€6bn pa (from ~€2bn pa currently) by 2035 with an associated capex of €8bn. We see some risk of cannibalization of some existing earnings	12	95.805496	Carbon	NEUT
Rio Tinto plc. FY'19 results in line, record dividend is 70% payout. Stating its case for being an energy transition leader. RIO LN. GPS-3279646-0	to 1.20 – 1.35/lb. Capexguidancewasreiteratedat7bn for 2020 and 6.5bnin2021/22.\r\n\r\n\r\nClimatechangeinitiativesatforefrontofstrategy 1bn of energy transition-related investment by 2025. RIO now : RIOintroducednew2030greenhousegasemissions(GHG)reductionstargetsandannouncedplansfor targets a 30% reduction in GHG emissions intensity by 2030 (vs 2018), plus a 15% absolute GHG emissions reduction by 2030 (vs 2018). They believe growth to 2030 can be carbon neutral. We have long regarded RIO as the Energy Transition leader among the Diversified Miners (link).\r\n\r\n\r\nAssessing its options: RIO indicated it has evaluated ~200 assets &	11	109.316383	Carbon	POS
Gazprom. Hydrogen: It's the future; but is there a role for Gazprom and turquoise in the hydrogen rainbow. OGZD LI. GPS-3390024-0	Turquoise H2 (gas/ CH4 pyrolysis) offers the potential for cheap H2 from gas but with zero process CO2 emissions. This is the technology BASF is working on and that Gazprom sees as the future for H2 from gas.\n\n\nGas and the hydrogen economy: Whilst gas may be viewed as a transition fuel by the industry, the transition period may yet be shorter than expected. The possibility of decarbonizing gas to hydrogen hence offers a potential next step in the evolution of gas as a fuel and to ensuring that gas continues to play a substantial role in the energy mix. For	11	112.895950	Carbon	NEUT
Alliance Resource Partners. Depressed Seaborne Prices Hurt H2 Outlook, but Case for Exports Remains Strong. ARLP US. GPS-2990016-0	+40-year lows in real terms. So while the move towards renewables will likely continue, the rate of switching from coal to gas (driven by very low gas prices) seems extended. It's impossible to know when US nat gas prices will normalize and what the true marginal gas cost will be – and to what extent international LNG prices will netback into the US gas market – but our confidence that gas prices are currently too low should support ARLP's longer-term coal operations, and it's also taking small steps into oil and gas to diversify its portfolio. Meanwhile the company continues	11	115.437592	Carbon	POS
EOG Resources. 1Q21 Preview: Anticipating Upside Financials and Cash Flow on Strong 1Q21 Realizations. EOG US. GPS-3700414-0	oil growth. A backwardated pricing environment would further reduce EOG's deferred ratio. We model $3.5 billion of FCF (pre-dividend) in 2021.\n\nCommitment to ESG: EOG has outlined its ESG ambitions, which include capturing 99.8% of wellhead gas in 2021 compared to its 99.6% capture rate in 2020 and reaching net zero scope 1 and 2 GHG emissions by 2040. EOG has also endorsed the World Bank Zero Routine Flaring by 2030 Initiative and targeting a 13.5 GHG emissions intensity rate and near-zero methane emissions percentage by 2025. The company also announced it will be expanding its closed-loop gas capture project in	11	111.890465	Carbon	POS
Kinder Morgan, Inc.. 2020 Investor Day Highlights Strict Adherence To Capital Discipline & Leveraging Secular NatGas Trends; Model Update. KMI US. GPS-3253847-0	for global demand. To quantify, by 2030 the non-OECD population will increase by ~850mm people to ~7.2bn, ~17x the growth of OECD countries. To attain a similar quality of life as OECD countries, non-OECD per capita energy demand would need to increase three times, pointing towards material incremental energy consumption. On the emissions front, the US and other OECD countries have achieved steady emissions reduction in recent years, with natural gas contributing to this success. As such, KMI sees growing energy demand from developing countries and lowering emissions through pivoting to natgas as supporting steady natgas demand growth. \r\r\n\r\r\n\r\r\nNatGas underpins	10	112.558853	Carbon	POS
The Williams Companies, Inc.. Model Update. WMB US. GPS-3623890-0	and long term energy goals, culminating in net zero emissions by 2050. Through 2025, WMB targets 3bnofcapexfocusedreducingemissions, includingaidingcoal – to – gasswitching, emissionsreductions, solarandRNGprojects. Ofthisportfolio, 35400mm. Hydrogen opportunities remain longer dated, but WMB notes its current infrastructure could blend hydrogen or eventually transition to full hydrogen. On the social performance side, WMB highlighted the CEO Action for Diversity & Inclusion Coalition, the D& Council, metrics dashboard, and D&I Training and Tools & Resources	10	100.752853	Carbon	NEUT
Tenneco Inc.. Reiterate Overweight Rating After Investor Tour Confirms Non-LV Sales and Margin Opportunity. TEN US. GPS-988606-0	production runs. TEN benefits from the need for greater fuel economy, not just cleaner emissions. We believe it is wrong to view emissions through a prism of "in-cylinder vs. aftertreatment" technologies. TEN noted the benefits of a holistic engine-fuel-exhaust approach to emissions control and fuel efficiency. The firm noted that some approaches to increasing fuel efficiency, such as higher combustion temperatures, can result in more NOx emissions. More effective aftertreatment, therefore, can allow an engine to burn hotter and more efficiently without concern of violating emission standards. In this way, TEN is levered to trend toward higher fuel economy as	10	103.653580	Carbon	POS

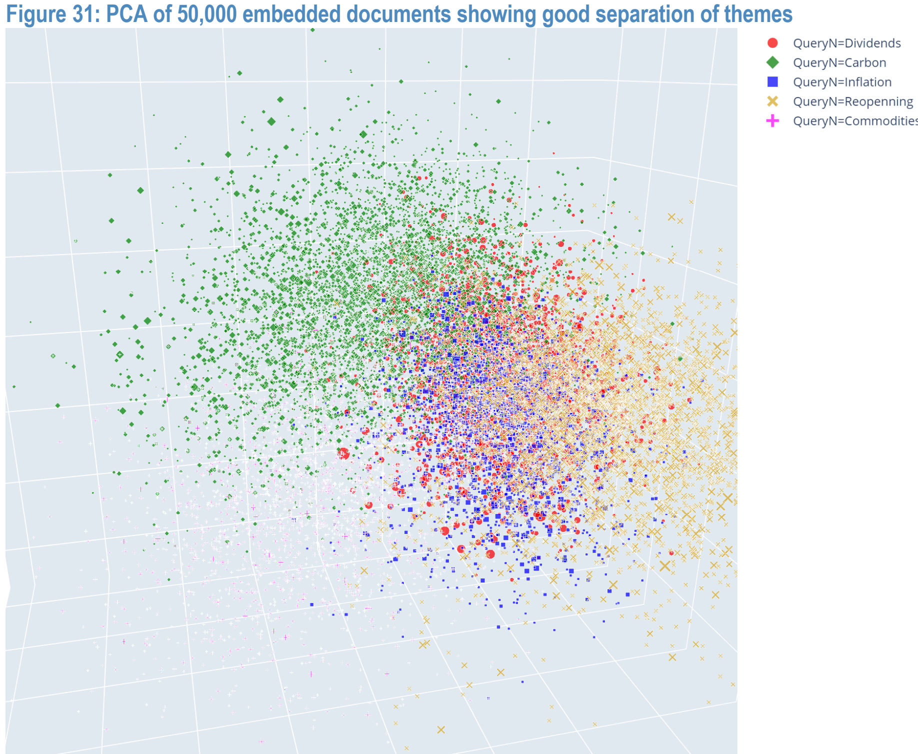
Source: J.P. Morgan QDS.

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

Visualisation of 1000's of Documents

We have used PCA analysis to extract the most significant 3 dimensions from our 748 embedding vectors to help with a visual representation. This example shows good separation of themes (clusters) here across 50,000 reports.

Figure 31: PCA of 50,000 embedded documents showing good separation of themes



Source: J.P. Morgan.

PyTorch for Large PCA

SKLearn struggles to perform PCA on large datasets, instead we find PyTorch to be amazingly quick. The use of torch helped with the visualization above.

Alternative Embedding Model

Alternative approach that looks promising is SPECTER: Document-level Representation Learning using Citation-informed Transformers by AllenAI and is also available on HuggingFace.

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

Not-Quite-Exact-Duplicate Removal

Often when scanning news articles, we find that there are multiple overlapping reports of the same event from different news vendors. Identifying and removing duplicate stories with non-exact matches can be difficult. However, with vector encoding of headlines, similar stories should have cosine distances close to zero and this can be used as a filter to help cluster and remove redundant articles.

We can use any vector model for this. The RAG model discussed above would be appropriate (and potentially faster for large volumes of news), while Spacy 3.0 also supports document vector encoding.

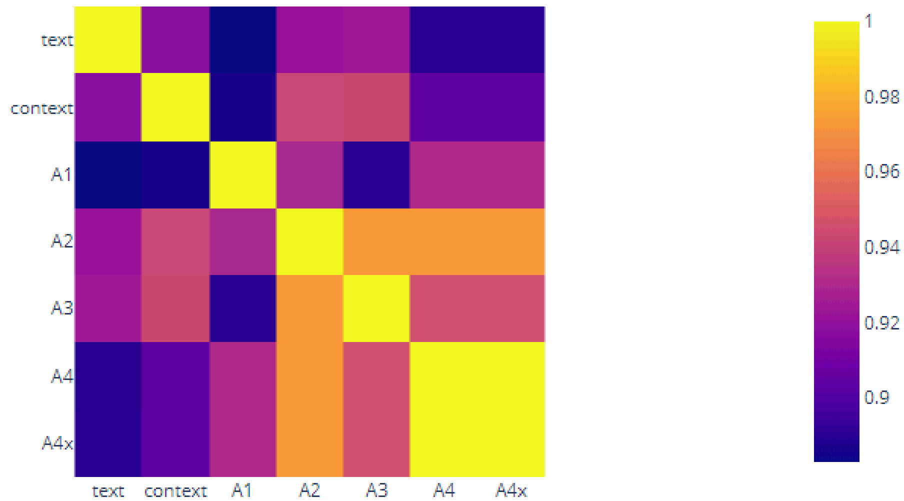
For this exercise we duplicated the last article (ARTICLE4) but reversed the word order and replaced 'year' with 'month' to test that a news article with different word order would have a similar vector encoding and cosine distance close to zero.

Figure 32: Sample Code to Calculate Similarity of Documents

```
import spacy
from tqdm.auto import tqdm
from sklearn.metrics import pairwise_distances
import plotly.express as px

# create a spacy nlp model
nlp = spacy.load('en_core_web_md')

# create an empty list to hold the vectors
vectors = [ ]

# We use TQDM to observe progress and time the batch job as we loop over a list of articles
for article in tqdm(articles):
    doc = nlp(article)
    vectors.append(doc.vector)

# Use sklearn's pariwise distance function to calculate cosine distance
cosine_similarity = 1-pairwise_distances(vectors, metric="cosine")
fig = px.imshow(cosine_similarity)
fig.show()
```

Source: Spacy, J.P. Morgan QDS.

Figure 33: Sample Matrix Showing Similarity of Documents



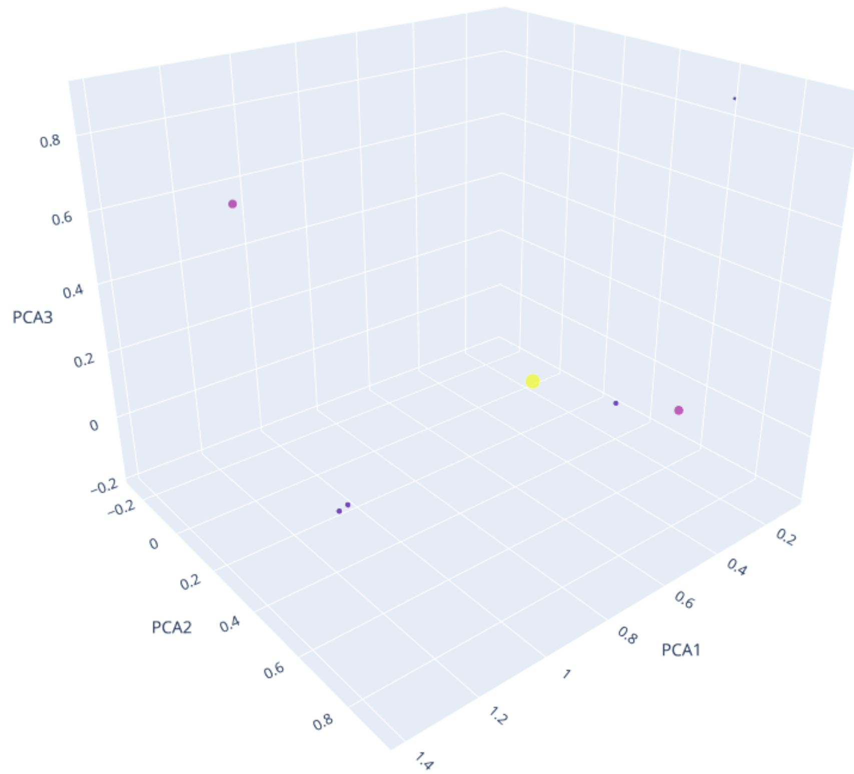
Source: Spacy, J.P. Morgan QDS.

Here we can easily identify the last two articles (A4 and A4x) as being duplicates, with similarity of 0.999 despite the fact that the word order was reversed, and a few words replaced.

We can also display the similarity of documents in 3-dimensional space after converting the 300 long vectors to just 3 principal components. The results below highlight the two similar articles in the foreground. The document length has been used to size & colour each dot.

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

Figure 34: Sample 3D PCA Analysis Showing Similarity of Documents by Proximity. Two nearly identical documents are shown as neighbors in the foreground.



Source: J.P. Morgan QDS.

Conclusions on Embedding Models

We consider embedding models as a 'lossy compression' technique. There are definite advantages to be had with these methods in reducing search time for relevant articles however this needs to be considered with the potential for inaccurate matches (false-positive and included but unwanted true-negative results). Another practical use of these vector models of documents is in clustering similar documents, and the identification (and potentially removal) of duplicates or near-match-duplicates.

NLP Pipelines

Using a mix of what we have been working with from various sources such as HuggingFace, FaceBook AI, Google Labs, Microsoft, and others, we have discovered a few useful tools for an efficient NLP pipeline production.

HuggingFace Transformers

Many NLP processing tasks can be handled by the transformers package. HuggingFace currently offer 17 pipelines to aid the use of NLP.

AudioClassificationPipeline	**SummarizationPipeline**
AutomaticSpeechRecognitionPipeline	**TableQuestionAnsweringPipeline**
ConversationalPipeline	**TextClassificationPipeline**
FeatureExtractionPipeline	**TextGenerationPipeline**
FillMaskPipeline	**Text2TextGenerationPipeline**
ImageClassificationPipeline	**TokenClassificationPipeline**
ImageSegmentationPipeline	**TranslationPipeline**
ObjectDetectionPipeline	**ZeroShotClassificationPipeline**
QuestionAnsweringPipeline	**SummarizationPipeline**

Spacy

Spacy is a popular and capable NLP tool. With the recent 3.1 release, support for various transformer models was added with the spacy-transformers package.

Datasets Package

Don't save data as CSV or other archaic formats. The Datasets package offers a fast and efficient dataframe style tool that supports *massive* datasets. When the data is larger than RAM, the package makes use of efficient memory mapping technology, essentially caching the dataset to disk automatically. To improve performance, chunks of processing can be applied using inbuilt multi-threading, with each row of data treated as a dict.

FAISS for index searching

FAISS was released in 2017 by Facebook as a library for efficient (exact or approximate) similarity search. This package is helpful for retrieving (approximate) nearest neighbors to a vector embedding search, very quickly, and can be extended to a large-scale networked compute solution if required.

Text Corpus (Data)

27 years of EDGAR filings are available at the Uni of Notre Dame. They also have a dictionary available which includes word level sentiment tags and frequency information which is often used in earlier NLP techniques such as TF-IDF. J.P. Morgan Research reports are available as RIXML articles as an additional service for interested clients.

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

NLP Compression for Faster Inference

In October 2021 we ran a DeepFin Investor Tutorial on Natural Language Modelling with NVIDIA held over video conference from London and New York. Part of the DeepFin session focused on 'Inference Mode' when the already-trained model can be streamlined to process answers faster. Before we make the model available either as a batch or streaming process for individuals to submit collected requests or interact with directly, we perform one final step to optimise BERT to perform inference. This is done using a NVIDIA library called TensorRT.

We optimise BERT to perform inference before making the model available to end-users.

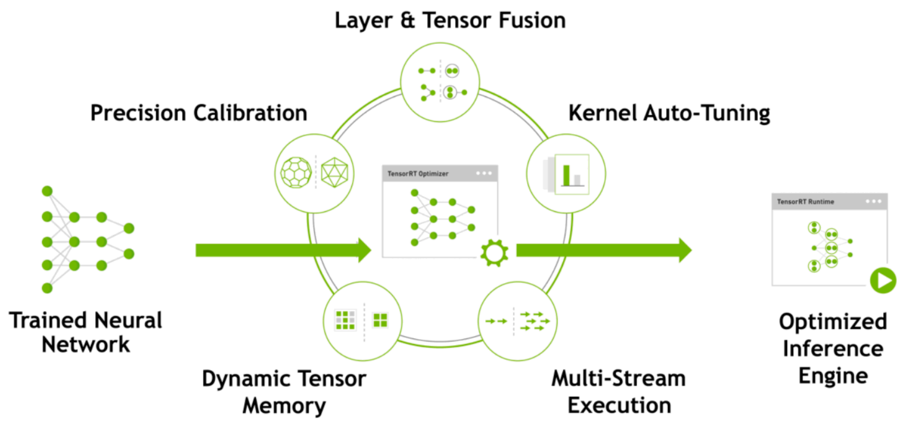
This is done using TensorRT a deep learning inference optimiser and runtime for low-latency and high-throughput inference applications.

See more at https://developer.nvidia.com/tensorrt

and

https://developer.nvidia.com/blog/nlu-with-tensorrt-bert/

Figure 35: TensorRT process and optimisations



Source: J.P. Morgan Quantitative and Derivatives Strategy, NVIDIA Inc.

TensorRT is an optimised inference engine for deep learning on GPUs. The training process is developed by framework libraries like TensorFlow (used in the DeepFin tutorial) and Pytorch. However, even though the inference forward pass is the same as what frameworks already do in training, we are now more interested in latency as the primary optimisation factor rather than throughput.

TensorRT ingests a trained model from any of the frameworks, builds a TensorRT version of that model that has been through both graph optimisations (flattening and pruning the DNN layers, reducing node count) and lower precision calibration (e.g. 16-bit instead of 32- or 64-bit accuracy) to decrease latency, and then finally is executed at runtime by TensorRT's low profile runtime engine. The DNN gradients that are vital for training the model are also deleted as they are of no further use for forward-only passes in inference mode. Latency of around 2ms is possible with the appropriate setup.

DeepSpeed

DeepSpeed package also helps improve speed of large models and helps avoid GPU memory errors. A large part of the improvements revolves around a novel solution, Zero Redundancy Optimizer (ZeRO), to optimize memory across multiple servers, vastly improving training speed while increasing the model size that can be efficiently trained. ZeRO has the potential to scale beyond 1 Trillion parameters using today's hardware. The researchers have used ZeRO to create one of the world's largest language model (Turing-NLG, 17B parameters as of 2020) with record breaking accuracy. The details on DeepSpeed improvements appear in the ZeRO and ZeRO2 papers.

Conclusions

We have shown a series of practical examples that make use of State of the Art (SotA) Natural Language Processing (NLP) and Artificial Intelligence (AI) tools to help with portfolio management tasks. The examples covered: Sentiment, Thematics, Q&A Summarisation & Embedding. We also offered some insights into building efficient NLP pipelines for text processing, clustering, charting and news article de-duplication.

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

Appendix 1: Long Text for Summarization

ARTICLE1 = """

New York (CNN) When Liana Barrientos was 23 years old, she got married in Westchester County, New York.

A year later, she got married again in Westchester County, but to a different man and without divorcing her first husband.

Only 18 days after that marriage, she got hitched yet again. Then, Barrientos declared "I do" five more times, sometimes only within two weeks of each other.

In 2010, she married once more, this time in the Bronx. In an application for a marriage license, she stated it was her "first and only" marriage.

Barrientos, now 39, is facing two criminal counts of "offering a false instrument for filing in the first degree," referring to her false statements on the

2010 marriage license application, according to court documents. Prosecutors said the marriages were part of an immigration scam.

On Friday, she pleaded not guilty at State Supreme Court in the Bronx, according to her attorney, Christopher Wright, who declined to comment further.

After leaving court, Barrientos was arrested and charged with theft of service and criminal trespass for allegedly sneaking into the New York subway through an emergency exit, said Detective

Annette Markowski, a police spokeswoman. In total, Barrientos has been married 10 times, with nine of her marriages occurring between 1999 and 2002.

All occurred either in Westchester County, Long Island, New Jersey or the Bronx. She is believed to still be married to four men, and at one time, she was married to eight men at once, prosecutors say.

Prosecutors said the immigration scam involved some of her husbands, who filed for permanent residence status shortly after the marriages.

Any divorces happened only after such filings were approved. It was unclear whether any of the men will be prosecuted.

The case was referred to the Bronx District Attorney\'s Office by Immigration and Customs Enforcement and the Department of Homeland Security\'s

Investigation Division. Seven of the men are from so-called "red-flagged" countries, including Egypt, Turkey, Georgia, Pakistan and Mali.

Her eighth husband, Rashid Rajput, was deported in 2006 to his native Pakistan after an investigation by the Joint Terrorism Task Force.

If convicted, Barrientos faces up to four years in prison. Her next court appearance is scheduled for May 18."""

Source: LexisNexis

ARTICLE2 = """Meet M6 — 10 Trillion Parameters at 1% GPT-3's Energy Cost

Smaller players can now enter the game of large AI models

Alberto Romero

Photo by vs148 on Shutterstock (edited)

I can confidently say artificial intelligence is advancing fast when a neural network 50 times larger than another can be trained at a 100 times less energy cost — with just one year in between!

On June 25, Alibaba DAMO Academy (the R&D branch of Alibaba) announced they had built M6, a large multimodal, multitasking language model with 1 trillion parameters — already 5x GPT-3's size, which serves as the standard to measure the rate of progress for large AI models. The model was intended for multimodality and multitasking, going a step further than previous models towards general intelligence.

In terms of abilities, M6 resembles GPT-3 and other similar models like Wu Dao 2.0 or MT-NGL 530B (from which we have very little information). InfoQ, a popular Chinese tech magazine compiles M6's main skills: "[It] has cognition and creativity beyond traditional AI, is good at drawing, writing, question and answer, and has broad application prospects in many fields such as e-commerce, manufacturing, literature and art."

However, the critical aspect Alibaba researchers highlighted was the significant efficiency and energy cost improvements. They reduce the consumption of the model by 80% and increased its efficiency x11 when compared to 100-million language models.

Extremely important news in line with green AI principles and objectives.

Green AI to demonopolize large language models

But they didn't stop there and now, 5 months later, they've just achieved not one, but two new striking milestones: They've improved M6 to make it the first 10-trillion-parameter large language model — 50x GPT-3's size. And they've bettered their previous marks on efficiency, reducing the energy consumption to 1% of what GPT-3 needed to train.

They used a mere 512 GPUs to train the model in 10 days!

These achievements will have far-reaching positive consequences for the AI community and the world.

On the one hand, it's a big leap towards finding common ground between the necessities of large AI models and the requirements of clean energy movements that aim at reducing the carbon footprint. One of the main criticisms of large language models is that they can't compensate for the huge amounts of pollution they generate. It's been estimated that training a large AI model (pre-GPT-3) contaminates 5 times more than a car in its entire lifetime — and their usefulness isn't so obvious. Amazon and Microsoft, among other tech companies, have already presented plans to reduce carbon emissions in the coming years, but both aim to tackle the problem by cooling the data centers whereas Alibaba has achieved a better solution; reducing the resources needed to train the models.

This has another important advantage. If Alibaba publishes the techniques and methods they've used to achieve its results, smaller players could enter into competition against the big tech corporations that are currently monopolizing the super-profitable field of large AI models.

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

The cost of researching, training, and inference creates such a toll that even giants like Google have had problems funding the technology. DeepMind, a Google subsidiary, decided not to investigate different possibilities for a key component when creating AlphaStar to avoid surpassing the budget.

OpenAI — which had access to a 10,000 Nvidia V100 supercomputer provided by Microsoft (although it hasn't been disclosed the exact amount of GPUs they used)— decided to not retrain GPT-3 after researchers found a mistake because it'd have been infeasible. Some gross calculations estimate a training cost of at least $4.6 million, which is out of reach for most companies — that's without including research and development costs, which would elevate the number to $10–30M.

How could smaller companies compete against that?

In contrast, the latest version of M6 has been trained on 512 GPUs for 10 days. (GPT-3 was trained on V100, but researchers calculated that using A100s, it would have taken 1,024 GPUs to train the model in 34 days.)

Doing some gross calculations we can compare the training cost for both models. I'll assume Alibaba used Nvidia A100 and a similar cost of GPU instance/hour as AWS, where an 8-Nvidia A100 AWS instance costs ~$20/hour. Given they used 512 GPUs, that makes 64 8-A100 instances. Doing the math we have the total cost = 64 #instances · $20/hour · 24 hours/day · 10 days = $307,200.

Still somewhat costly, but nowhere near what OpenAI spent to train GPT-3.

A silver lining for the future

In the past, I've been very critical of large language models for reasons ranging from discrimination and biases to capacity for misinformation, to lack of understanding, and even because why do we even need more large language models? And also because of the high environmental and financial costs creating these systems entails.

But today I applaud the results Alibaba DAMO Academy has published.

It seems they're committed to improving at least some of the problems this new AI trend carries. There's still a lot of work to do — and some of the issues are so intrinsic to these models that we can only hope to mitigate them — but seeing big tech companies aiming to improve the current landscape is a silver lining for the near-term future of artificial intelligence.

If you liked this article, consider subscribing to my free weekly newsletter Minds of Tomorrow! News, research, and insights on Artificial Intelligence every week!

You can also support my work directly and get unlimited access by becoming a Medium member using my referral link here! :)

Every Thursday, the Variable delivers the very best of Towards Data Science: from hands-on tutorials and cutting-edge research to original features you don't want to miss. Take a look.

""

Source: LexisNexis

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

ARTICLE3 = """

Wind power experienced a substantial growth over the past decade especially because it has been seen as one of the best ways towards meeting climate change and emissions targets by many countries.

Since wind power is not fully dispatchable, the accuracy of wind forecasts is a key element for the electric system operators, as it strongly affects the decision-making processes.

The planning horizon can be short term (1 -3 months) and long-term (6–12 months) depending on the process.

The objective of this paper is to conduct a performance comparison of five deep learning models each combined with three types of data pre-processing and used for short term and long-term multi-variate predictions.

The input data are time series of the wind power capacity factor and the temperature. In addition, this paper sets out to demonstrate and review the state-of-the-art deep learning models for prediction

with a secondary objective to present the reader a reference point to better understand which model to choose and what factors are significant.

The first contribution of this paper is to apply, assess and compare a selection of the novel and cutting-edge deep learning models for multi-variate prediction.

Multi-variate predication is achieved through a proposed multiple input and multiple output (MIMO) architecture. Compared to traditional prediction models, machine learning techniques have the advantage of generalization.

Among various techniques deep learning is particularly getting more attention due to the applicability to various dataset such as numerical and character.

This investigation focuses on five models — Deep Feed Forward (DFF), Deep Convolutional Network (DCN), Recurrent Neural Network (RNN), Attention mechanism (Attention) and Long Short-Term Memory Networks (LSTM).

The second contribution is to propose a novel approach to transform the time series dataset to signal for input and reconstruct the model predictions through inverse transformation,

by means of the so-called discrete wavelet transformation and fast Fourier transformation. The different models are assessed also by comparing their performance with and

without the input dataset manipulation through wavelet and FFT transformation. Beyond that, the model performances are outlined in detail, to give the reader an overview of the models to choose

from for short-term or long-term prediction. The results demonstrate that the Attention and DCN perform best with Wavelet or FFT signal, whereas some other models perform better with no data preprocessing."""

Source: LexisNexis

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

ARTICLE4 = """

Our underlying profit of $447 million was smaller than usual but was still a very sound result, given the scale of the challenges. If you recall a year ago, our industry was dealing with the shockwaves arising from the confluence of demand contraction from the pandemic and a price war between OPEC and Russia that resulted in global oversupply of oil.

Our reported loss, of just over $4 billion, reflects the major writedowns of our assets announced in July. At a time of extreme volatility and uncertainty in global markets, we took a prudent view on the carrying value of those assets.

I know that 2020 was also a tough year for shareholders as our share price reflected the difficult environment I have spoken about, and even though we were able to pay dividends, they were consequently less than in previous years.

2021 Annual General Meeting

"""

Source: Woodside Company Report

Berowne Hlavaty
(61-2) 9003-8602
berowne.d.hlavaty@jpmorgan.com

Appendix 2: Additional References

Interesting overview of Facebook's NLP techniques and a comparison of Q&A methods.

Retrieval-Augmented Generation (RAG) Models look for answers similar to a question. How to Build a RAG, alternatively with RayRag or Haystack by deepset.

The *Fusion-in-Decoder* approach, proposed by Izacard & Grave (2020) is also based on a pre-trained T5 model.

An example on medium shows how to use coreferences to build a relationship graph.

Another paper looks at "Improving Multi-hop Question Answering over Knowledge Graphs using Knowledge Base Embeddings".

NLP can be used to find quoted text from the Guardian.

Mixed Mode learning is also becoming prevalent, such as models that can perform Q&A with video input, like this one called Violet.

Finding the shortest dependency path for NLP discussed on Towards Data Science.

Multiple-element joint detection for Aspect-Based Sentiment Analysis

Cross-Domain End-To-End Aspect-Based Sentiment Analysis with Domain-Dependent Embeddings

Github implementation of EMNLP 2020 paper titled GRACE: Gradient Harmonized and Cascaded Labeling for Aspect-based Sentiment Analysis.

10 Leading Language Models for NLP [in 2019 to 2020]

Aspect-Controllable Opinion Summarization looks like a potential option for pre-processing before sentiment prediction.

Analyst Certification: The Research Analyst(s) denoted by an “AC” on the cover of this report certifies (or, where multiple Research Analysts are primarily responsible for this report, the Research Analyst denoted by an “AC” on the cover or within the document individually certifies, with respect to each security or issuer that the Research Analyst covers in this research) that: (1) all of the views expressed in this report accurately reflect the Research Analyst’s personal views about any and all of the subject securities or issuers; and (2) no part of any of the Research Analyst's compensation was, is, or will be directly or indirectly related to the specific recommendations or views expressed by the Research Analyst(s) in this report. For all Korea-based Research Analysts listed on the front cover, if applicable, they also certify, as per KOFIA requirements, that the Research Analyst’s analysis was made in good faith and that the views reflect the Research Analyst’s own opinion, without undue influence or intervention.

All authors named within this report are Research Analysts unless otherwise specified. In Europe, Sector Specialists (Sales and Trading) may be shown on this report as contacts but are not authors of the report or part of the Research Department.

Research excerpts: This material may include excerpts from previously published research. For access to the full reports, including analyst certification and important disclosures, investment thesis, valuation methodology, and risks to rating and price targets, please contact your salesperson or the covering analyst’s team or visit www.jpmorganmarkets.com.

Important Disclosures

This report is a product of the research department's Global Quantitative and Derivatives Strategy group. Views expressed may differ from the views of the research analysts covering stocks or sectors mentioned in this report. Structured securities, options, futures and other derivatives are complex instruments, may involve a high degree of risk, and may be appropriate investments only for sophisticated investors who are capable of understanding and assuming the risks involved. Because of the importance of tax considerations to many option transactions, investors considering options should consult with their tax advisor as to how taxes affect the outcome of contemplated option transactions.

Company-Specific Disclosures: Important disclosures, including price charts and credit opinion history tables, are available for compendium reports and all J.P. Morgan–covered companies, and certain non-covered companies, by visiting https://www.jpmm.com/research/disclosures, calling 1-800-477-0406, or e-mailing research.disclosure.inquiries@jpmorgan.com with your request.

Explanation of Equity Research Ratings, Designations and Analyst(s) Coverage Universe:
J.P. Morgan uses the following rating system: Overweight [Over the next six to twelve months, we expect this stock will outperform the average total return of the stocks in the analyst’s (or the analyst’s team’s) coverage universe.] Neutral [Over the next six to twelve months, we expect this stock will perform in line with the average total return of the stocks in the analyst’s (or the analyst’s team’s) coverage universe.] Underweight [Over the next six to twelve months, we expect this stock will underperform the average total return of the stocks in the analyst’s (or the analyst’s team’s) coverage universe.] Not Rated (NR): J.P. Morgan has removed the rating and, if applicable, the price target, for this stock because of either a lack of a sufficient fundamental basis or for legal, regulatory or policy reasons. The previous rating and, if applicable, the price target, no longer should be relied upon. An NR designation is not a recommendation or a rating. In our Asia (ex-Australia and ex-India) and U.K. small- and mid-cap equity research, each stock’s expected total return is compared to the expected total return of a benchmark country market index, not to those analysts’ coverage universe. If it does not appear in the Important Disclosures section of this report, the certifying analyst’s coverage universe can be found on J.P. Morgan’s research website, www.jpmorganmarkets.com.

J.P. Morgan Equity Research Ratings Distribution, as of January 01, 2022

	Overweight (buy)	**Neutral** (hold)	**Underweight** (sell)
J.P. Morgan Global Equity Research Coverage*	52%	37%	11%
IB clients**	53%	46%	34%
JPMS Equity Research Coverage*	51%	37%	12%
IB clients**	74%	68%	50%

*Please note that the percentages might not add to 100% because of rounding.
**Percentage of subject companies within each of the "buy," "hold" and "sell" categories for which J.P. Morgan has provided investment banking services within the previous 12 months.
For purposes only of FINRA ratings distribution rules, our Overweight rating falls into a buy rating category; our Neutral rating falls into a hold rating category; and our Underweight rating falls into a sell rating category. Please note that stocks with an NR designation are not included in the table above. This information is current as of the end of the most recent calendar quarter.

Equity Valuation and Risks: For valuation methodology and risks associated with covered companies or price targets for covered companies, please see the most recent company-specific research report at http://www.jpmorganmarkets.com, contact the primary analyst or your J.P. Morgan representative, or email research.disclosure.inquiries@jpmorgan.com. For material information about the proprietary models used, please see the Summary of Financials in company-specific research reports and the Company Tearsheets, which are available to download on the company pages of our client website, http://www.jpmorganmarkets.com. This report also sets out within it the material underlying assumptions used.

A history of J.P. Morgan investment recommendations disseminated during the preceding 12 months can be accessed on the Research & Commentary page of http://www.jpmorganmarkets.com where you can also search by analyst name, sector or financial instrument.

Analysts' Compensation: The research analysts responsible for the preparation of this report receive compensation based upon various factors, including the quality and accuracy of research, client feedback, competitive factors, and overall firm revenues.

Registration of non-US Analysts: Unless otherwise noted, the non-US analysts listed on the front of this report are employees of non-US affiliates of J.P. Morgan Securities LLC, may not be registered as research analysts under FINRA rules, may not be associated persons of J.P. Morgan Securities LLC, and may not be subject to FINRA Rule 2241 or 2242 restrictions on communications with covered companies, public appearances, and trading securities held by a research analyst account.

Other Disclosures

J.P. Morgan is a marketing name for investment banking businesses of JPMorgan Chase & Co. and its subsidiaries and affiliates worldwide.

All research material made available to clients are simultaneously available on our client website, J.P. Morgan Markets, unless specifically permitted by relevant laws. Not all research content is redistributed, e-mailed or made available to third-party aggregators. For all research material available on a particular stock, please contact your sales representative.

Any long form nomenclature for references to China; Hong Kong; Taiwan; and Macau within this research material are Mainland China; Hong Kong SAR (China); Taiwan (China); and Macau SAR (China).

J.P. Morgan Research may, from time to time, write on issuers or securities targeted by economic or financial sanctions imposed or administered by the governmental authorities of the U.S., EU, UK or other relevant jurisdictions (Sanctioned Securities). Nothing in this report is intended to be read or construed as encouraging, facilitating, promoting or otherwise approving investment or dealing in such Sanctioned Securities. Clients should be aware of their own legal and compliance obligations when making investment decisions.

Options and Futures related research: If the information contained herein regards options- or futures-related research, such information is available only to persons who have received the proper options or futures risk disclosure documents. Please contact your J.P. Morgan Representative or visit https://www.theocc.com/components/docs/riskstoc.pdf for a copy of the Option Clearing Corporation's Characteristics and Risks of Standardized Options or http://www.finra.org/sites/default/files/Security_Futures_Risk_Disclosure_Statement_2018.pdf for a copy of the Security Futures Risk Disclosure Statement.

Changes to Interbank Offered Rates (IBORs) and other benchmark rates: Certain interest rate benchmarks are, or may in the future become, subject to ongoing international, national and other regulatory guidance, reform and proposals for reform. For more information, please consult: https://www.jpmorgan.com/global/disclosures/interbank_offered_rates

Private Bank Clients: Where you are receiving research as a client of the private banking businesses offered by JPMorgan Chase & Co. and its subsidiaries ("J.P. Morgan Private Bank"), research is provided to you by J.P. Morgan Private Bank and not by any other division of J.P. Morgan, including, but not limited to, the J.P. Morgan Corporate and Investment Bank and its Global Research division.

Legal entity responsible for the production and distribution of research: The legal entity identified below the name of the Reg AC Research Analyst who authored this material is the legal entity responsible for the production of this research. Where multiple Reg AC Research Analysts authored this material with different legal entities identified below their names, these legal entities are jointly responsible for the production of this research. Research Analysts from various J.P. Morgan affiliates may have contributed to the production of this material but may not be licensed to carry out regulated activities in your jurisdiction (and do not hold themselves out as being able to do so). Unless otherwise stated below, this material has been distributed by the legal entity responsible for production. If you have any queries, please contact the relevant Research Analyst in your jurisdiction or the entity in your jurisdiction that has distributed this research material.

Legal Entities Disclosures and Country-/Region-Specific Disclosures:
Argentina: JPMorgan Chase Bank N.A Sucursal Buenos Aires is regulated by Banco Central de la República Argentina ("BCRA"- Central Bank of Argentina) and Comisión Nacional de Valores ("CNV"- Argentinian Securities Commission" - ALYC y AN Integral N°51). **Australia:** J.P. Morgan Securities Australia Limited ("JPMSAL") (ABN 61 003 245 234/AFS Licence No: 238066) is regulated by the Australian Securities and Investments Commission and is a Market, Clearing and Settlement Participant of ASX Limited and CHI-X. This material is issued and distributed in Australia by or on behalf of JPMSAL only to "wholesale clients" (as defined in section 761G

of the Corporations Act 2001). A list of all financial products covered can be found by visiting https://www.jpmm.com/research/disclosures. J.P. Morgan seeks to cover companies of relevance to the domestic and international investor base across all Global Industry Classification Standard (GICS) sectors, as well as across a range of market capitalisation sizes. If applicable, in the course of conducting public side due diligence on the subject company(ies), the Research Analyst team may at times perform such diligence through corporate engagements such as site visits, discussions with company representatives, management presentations, etc. Research issued by JPMSAL has been prepared in accordance with J.P. Morgan Australia's Research Independence Policy which can be found at the following link: J.P. Morgan Australia - Research Independence Policy. **Brazil**: Banco J.P. Morgan S.A. is regulated by the Comissao de Valores Mobiliarios (CVM) and by the Central Bank of Brazil. Ombudsman J.P. Morgan: 0800-7700847 / ouvidoria.jp.morgan@jpmorgan.com. **Canada**: J.P. Morgan Securities Canada Inc. is a registered investment dealer, regulated by the Investment Industry Regulatory Organization of Canada and the Ontario Securities Commission and is the participating member on Canadian exchanges. This material is distributed in Canada by or on behalf of J.P.Morgan Securities Canada Inc. **Chile:** Inversiones J.P. Morgan Limitada is an unregulated entity incorporated in Chile. **China:** J.P. Morgan Securities (China) Company Limited has been approved by CSRC to conduct the securities investment consultancy business. **Dubai International Financial Centre (DIFC)**: JPMorgan Chase Bank, N.A., Dubai Branch is regulated by the Dubai Financial Services Authority (DFSA) and its registered address is Dubai International Financial Centre - The Gate, West Wing, Level 3 and 9 PO Box 506551, Dubai, UAE. This material has been distributed by JP Morgan Chase Bank, N.A., Dubai Branch to persons regarded as professional clients or market counterparties as defined under the DFSA rules. **European Economic Area (EEA):** Unless specified to the contrary, research is distributed in the EEA by J.P. Morgan AG ("JPM AG"), which is a member of the Frankfurt Stock Exchange, is authorised by the European Central Bank ("ECB") and is regulated by the Federal Financial Supervisory Authority (BaFin). JPM AG is a company incorporated in the Federal Republic of Germany with a registered office at Taunustor 1, 60310 Frankfurt am Main, the Federal Republic of Germany. The material has been distributed in the EEA to persons regarded as professional investors (or equivalent) pursuant to Art. 4 para. 1 no. 10 and Annex II of MiFID II and its respective implementation in their home jurisdictions ("EEA professional investors"). This material must not be acted on or relied on by persons who are not EEA professional investors. Any investment or investment activity to which this material relates is only available to EEA relevant persons and will be engaged in only with EEA relevant persons. **Hong Kong**: J.P. Morgan Securities (Asia Pacific) Limited (CE number AAJ321) is regulated by the Hong Kong Monetary Authority and the Securities and Futures Commission in Hong Kong, and J.P. Morgan Broking (Hong Kong) Limited (CE number AAB027) is regulated by the Securities and Futures Commission in Hong Kong. JP Morgan Chase Bank, N.A., Hong Kong (CE Number AAL996) is regulated by the Hong Kong Monetary Authority and the Securities and Futures Commission, is organized under the laws of the United States with limited liability. **India**: J.P. Morgan India Private Limited (Corporate Identity Number - U67120MH1992FTC068724), having its registered office at J.P. Morgan Tower, Off. C.S.T. Road, Kalina, Santacruz - East, Mumbai – 400098, is registered with the Securities and Exchange Board of India (SEBI) as a 'Research Analyst' having registration number INH000001873. J.P. Morgan India Private Limited is also registered with SEBI as a member of the National Stock Exchange of India Limited and the Bombay Stock Exchange Limited (SEBI Registration Number – INZ000239730) and as a Merchant Banker (SEBI Registration Number - MB/INM000002970). Telephone: 91-22-6157 3000, Facsimile: 91-22-6157 3990 and Website: www.jpmipl.com. JPMorgan Chase Bank, N.A. - Mumbai Branch is licensed by the Reserve Bank of India (RBI) (Licence No. 53/ Licence No. BY.4/94; SEBI - IN/CUS/014/ CDSL : IN-DP-CDSL-444-2008/ IN-DP-NSDL-285-2008/ INBI00000984/ INE231311239) as a Scheduled Commercial Bank in India, which is its primary license allowing it to carry on Banking business in India and other activities, which a Bank branch in India are permitted to undertake. For non-local research material, this material is not distributed in India by J.P. Morgan India Private Limited. **Indonesia**: PT J.P. Morgan Sekuritas Indonesia is a member of the Indonesia Stock Exchange and is regulated by the OJK a.k.a. BAPEPAM LK. **Korea**: J.P. Morgan Securities (Far East) Limited, Seoul Branch, is a member of the Korea Exchange (KRX). JPMorgan Chase Bank, N.A., Seoul Branch, is licensed as a branch office of foreign bank (JPMorgan Chase Bank, N.A.) in Korea. Both entities are regulated by the Financial Services Commission (FSC) and the Financial Supervisory Service (FSS). For non-macro research material, the material is distributed in Korea by or through J.P. Morgan Securities (Far East) Limited, Seoul Branch. **Japan**: JPMorgan Securities Japan Co., Ltd. and JPMorgan Chase Bank, N.A., Tokyo Branch are regulated by the Financial Services Agency in Japan. **Malaysia**: This material is issued and distributed in Malaysia by JPMorgan Securities (Malaysia) Sdn Bhd (18146-X), which is a Participating Organization of Bursa Malaysia Berhad and holds a Capital Markets Services License issued by the Securities Commission in Malaysia. **Mexico**: J.P. Morgan Casa de Bolsa, S.A. de C.V.and J.P. Morgan Grupo Financiero are members of the Mexican Stock Exchange and are authorized to act as a broker dealer by the National Banking and Securities Exchange Commission. **New Zealand**: This material is issued and distributed by JPMSAL in New Zealand only to "wholesale clients" (as defined in the Financial Advisers Act 2008). JPMSAL is registered as a Financial Service Provider under the Financial Service providers (Registration and Dispute Resolution) Act of 2008. **Pakistan**: J. P. Morgan Pakistan Broking (Pvt.) Ltd is a member of the Karachi Stock Exchange and regulated by the Securities and Exchange Commission of Pakistan. **Philippines**: J.P. Morgan Securities Philippines Inc. is a Trading Participant of the Philippine Stock Exchange and a member of the Securities Clearing Corporation of the Philippines and the Securities Investor Protection Fund. It is regulated by the Securities and Exchange Commission. **Russia**: CB J.P. Morgan Bank International LLC is regulated by the Central Bank of Russia. **Singapore**: This material is issued and distributed in Singapore by or through J.P. Morgan Securities Singapore Private Limited (JPMSS) [MCI (P) 093/09/2021 and Co. Reg. No.: 199405335R], which is a member of the Singapore Exchange Securities Trading Limited, and/or JPMorgan Chase Bank, N.A., Singapore branch (JPMCB Singapore), both of which are regulated by the Monetary Authority of Singapore. This material is issued and distributed in Singapore only to accredited investors, expert investors and institutional investors, as defined in Section 4A of the Securities and Futures Act, Cap. 289 (SFA). This material is not intended to be issued or distributed to any retail investors or any other investors that do not fall into the classes of "accredited investors," "expert investors" or "institutional investors," as defined under Section 4A of the SFA.

Recipients of this material in Singapore are to contact JPMSS or JPMCB Singapore in respect of any matters arising from, or in connection with, the material. As at the date of this material, JPMSS is a designated market maker for certain structured warrants listed on the Singapore Exchange where the underlying securities may be the securities discussed in this material. Arising from its role as a designated market maker for such structured warrants, JPMSS may conduct hedging activities in respect of such underlying securities and hold or have an interest in such underlying securities as a result. The updated list of structured warrants for which JPMSS acts as designated market maker may be found on the website of the Singapore Exchange Limited: http://www.sgx.com. **South Africa**: J.P. Morgan Equities South Africa Proprietary Limited and JPMorgan Chase Bank, N.A., Johannesburg Branch are members of the Johannesburg Securities Exchange and are regulated by the Financial Services Board. **Taiwan**: J.P. Morgan Securities (Taiwan) Limited is a participant of the Taiwan Stock Exchange (company-type) and regulated by the Taiwan Securities and Futures Bureau. Material relating to equity securities is issued and distributed in Taiwan by J.P. Morgan Securities (Taiwan) Limited, subject to the license scope and the applicable laws and the regulations in Taiwan. According to Paragraph 2, Article 7-1 of Operational Regulations Governing Securities Firms Recommending Trades in Securities to Customers (as amended or supplemented) and/or other applicable laws or regulations, please note that the recipient of this material is not permitted to engage in any activities in connection with the material that may give rise to conflicts of interests, unless otherwise disclosed in the "Important Disclosures" in this material. **Thailand**: This material is issued and distributed in Thailand by JPMorgan Securities (Thailand) Ltd., which is a member of the Stock Exchange of Thailand and is regulated by the Ministry of Finance and the Securities and Exchange Commission, and its registered address is 3rd Floor, 20 North Sathorn Road, Silom, Bangrak, Bangkok 10500. **UK:** Unless specified to the contrary, research is distributed in the UK by J.P. Morgan Securities plc ("JPMS plc") which is a member of the London Stock Exchange and is authorised by the Prudential Regulation Authority and regulated by the Financial Conduct Authority and the Prudential Regulation Authority. JPMS plc is registered in England & Wales No. 2711006, Registered Office 25 Bank Street, London, E14 5JP. This material is directed in the UK only to: (a) persons having professional experience in matters relating to investments falling within article 19(5) of the Financial Services and Markets Act 2000 (Financial Promotion) (Order) 2005 ("the FPO"); (b) persons outlined in article 49 of the FPO (high net worth companies, unincorporated associations or partnerships, the trustees of high value trusts, etc.); or (c) any persons to whom this communication may otherwise lawfully be made; all such persons being referred to as "UK relevant persons". This material must not be acted on or relied on by persons who are not UK relevant persons. Any investment or investment activity to which this material relates is only available to UK relevant persons and will be engaged in only with UK relevant persons. Research issued by JPMS plc has been prepared in accordance with JPMS plc's policy for prevention and avoidance of conflicts of interest related to the production of Research which can be found at the following link: J.P. Morgan EMEA - Research Independence Policy. **U.S.**: J.P. Morgan Securities LLC ("JPMS") is a member of the NYSE, FINRA, SIPC, and the NFA. JPMorgan Chase Bank, N.A. is a member of the FDIC. Material published by non-U.S. affiliates is distributed in the U.S. by JPMS who accepts responsibility for its content.

General: Additional information is available upon request. The information in this material has been obtained from sources believed to be reliable. While all reasonable care has been taken to ensure that the facts stated in this material are accurate and that the forecasts, opinions and expectations contained herein are fair and reasonable, JPMorgan Chase & Co. or its affiliates and/or subsidiaries (collectively J.P. Morgan) make no representations or warranties whatsoever to the completeness or accuracy of the material provided, except with respect to any disclosures relative to J.P. Morgan and the Research Analyst's involvement with the issuer that is the subject of the material. Accordingly, no reliance should be placed on the accuracy, fairness or completeness of the information contained in this material. Any data discrepancies in this material could be the result of different calculations and/or adjustments. J.P. Morgan accepts no liability whatsoever for any loss arising from any use of this material or its contents, and neither J.P. Morgan nor any of its respective directors, officers or employees, shall be in any way responsible for the contents hereof, apart from the liabilities and responsibilities that may be imposed on them by the relevant regulatory authority in the jurisdiction in question, or the regulatory regime thereunder. Opinions, forecasts or projections contained in this material represent J.P. Morgan's current opinions or judgment as of the date of the material only and are therefore subject to change without notice. Periodic updates may be provided on companies/industries based on company-specific developments or announcements, market conditions or any other publicly available information. There can be no assurance that future results or events will be consistent with any such opinions, forecasts or projections, which represent only one possible outcome. Furthermore, such opinions, forecasts or projections are subject to certain risks, uncertainties and assumptions that have not been verified, and future actual results or events could differ materially. The value of, or income from, any investments referred to in this material may fluctuate and/or be affected by changes in exchange rates. All pricing is indicative as of the close of market for the securities discussed, unless otherwise stated. Past performance is not indicative of future results. Accordingly, investors may receive back less than originally invested. This material is not intended as an offer or solicitation for the purchase or sale of any financial instrument. The opinions and recommendations herein do not take into account individual client circumstances, objectives, or needs and are not intended as recommendations of particular securities, financial instruments or strategies to particular clients. The recipients of this material must make their own independent decisions regarding any securities or financial instruments mentioned herein and should seek advice from such independent financial, legal, tax or other adviser as they deem necessary. J.P. Morgan may trade as a principal on the basis of the Research Analysts' views and research, and it may also engage in transactions for its own account or for its clients' accounts in a manner inconsistent with the views taken in this material, and J.P. Morgan is under no obligation to ensure that such other communication is brought to the attention of any recipient of this material. Others within J.P. Morgan, including Strategists, Sales staff and other Research Analysts, may take views that are inconsistent with those taken in this material. Employees of J.P. Morgan not involved in the preparation of this material may have investments in the securities (or derivatives of such securities) mentioned in this material and may trade them in ways different from those discussed in this material. This material is not an advertisement for or marketing of any issuer, its products or services, or its securities in any jurisdiction.

"Other Disclosures" last revised January 01, 2022.