4. UNLOCKING INSIGHTS AND OPPORTUNITIES WITH NLP IN ASSET MANAGEMENT

Andrew Chin
Head of Quantitative Research and Chief Data Scientist, AllianceBernstein

Yuyu Fan
Senior Data Scientist, AllianceBernstein

Che Guan
Senior Data Scientist, AllianceBernstein

Introduction

A confluence of events is affecting the asset management industry, forcing industry participants to rethink their competitive positioning and evolve to survive in the new world order. Geopolitical, regulatory, technological, and social trends are upending long-held norms, and the status quo is likely an untenable option for many firms. These forces are creating a host of challenges for existing players and presenting new opportunities for emerging firms. We discuss some of the key challenges affecting the active management industry in this new environment. While our list is not meant to be exhaustive, we focus on the main trends that will drive the adoption of text mining techniques in the coming years. We also provide the motivation for firms to leverage natural language processing (NLP) to capitalize on these trends.

Low Expected Returns

The driving focus for many investors since the Global Financial Crisis (GFC) has been the search for returns. Bond yields collapsed following the GFC, and concerns about economic growth suppressed expectations around equity returns. With prospects for returns expected to be much lower versus historical norms, asset owners and asset managers widened their search for high-yielding and high-returning assets. At the end of 2021, US 10-year government yields hovered near all-time lows, at 1.5%, and the price-to-earnings ratio of the S&P 500 Index was at 25, higher than historical averages. Although inflationary concerns caused a rise in rates and a significant drawdown in equity markets over the first nine months of 2022, below-average yields and above-average equity valuations persist across many countries, implying that future returns will likely be on the low end of long-term trends.

Over the past decade, investors are increasingly taking on more risk in an effort to enhance returns. Private markets and structured products are becoming more popular in the asset allocations of many investors. While these historically niche areas are becoming more mainstream, the investment, operational, and counterparty risks associated with them may still not be fully understood. Nevertheless, the low expected returns in the current environment are pushing asset managers to find new sources of returns and differentiation.

Active Managers Have Struggled

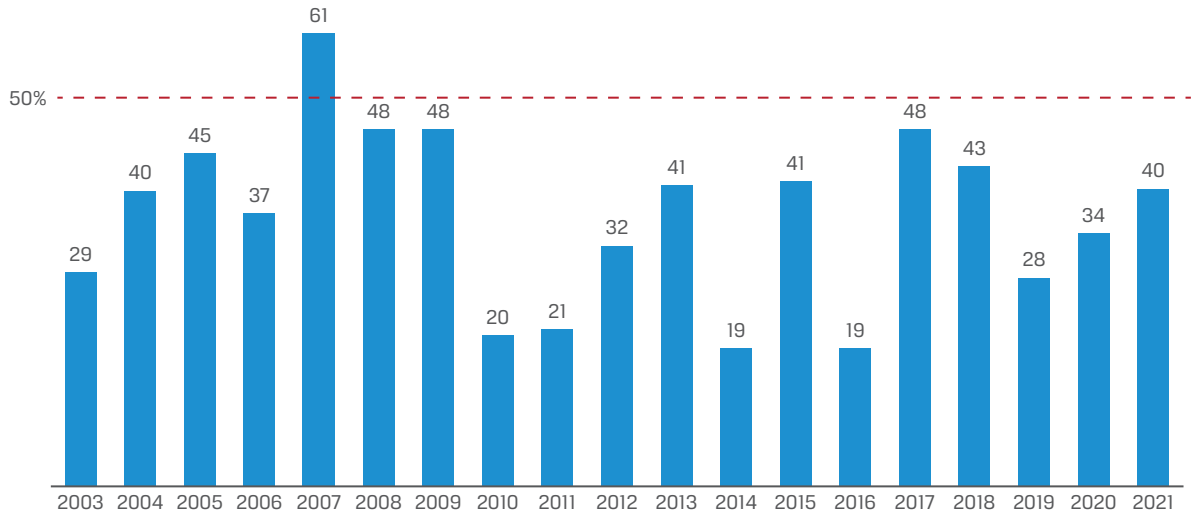
While investors are searching for higher returns, active managers, overall, have not delivered on their promises. Over the past decade, active managers have underperformed their respective benchmarks. **Exhibit 1** shows the percentage of US large-cap equity managers outperforming the broad US market as measured by the Russell 1000 Index. During any given year since the GFC, about one-third of the managers have beaten their benchmarks, suggesting that over longer horizons, even fewer managers are consistently beating the markets and providing value for their clients.

As a result of these struggles, management fees have collapsed, and significant assets have moved from high-fee, active to low-fee, passive strategies. Specifically, index products at the largest asset managers have swelled over the past decade, adding further to the headwinds for active management. One only needs to witness the enormous growth of such popular products as the SPDR S&P 500 ETF Trust (SPY) or the Vanguard 500 Index ETF (VOO) to gauge investor preferences. Many believe that these trends are set to persist unless active managers can reverse their recent headwinds.

Big Data and Data Science Present Opportunities

Given these challenges, asset managers are looking to provide higher and more consistent returns for their clients. For long-only managers, market returns dominate total portfolio returns; most of these managers have a beta close to 1 against their benchmark, and as a result, portfolio returns will largely mimic the returns of the broader market. Managers may look outside their stated

Exhibit 1. Percentage of Funds Outperforming the Russell 1000 Index, 2003–2021



Source: Bank of America.

benchmarks to enhance their returns. For example, equity managers may include IPOs (initial public offerings) and SPACs (special purpose acquisition companies) to potentially increase returns, while bond managers may include securitized instruments or other higher-yielding assets to enhance performance. All these strategies look "outside the benchmark" for investment opportunities and attempt to enhance the total returns of the portfolios.

Managers look to provide a consistent edge above and beyond the broad markets by attempting to uncover differentiated insights around potential investments. They endeavor to gain a deeper understanding of their investments, giving them more confidence in the companies or the securities they are interested in. Examples include a consumer analyst having a better sense of customer preferences for a certain brand or a tech analyst being able to forecast the technology stack favored by companies in the future. Other examples may include a systematic process that leverages unique data sources or uses sophisticated algorithms to synthesize data. These insights can give organizations the confidence they need to bolster their convictions and deliver stronger performance for their clients.

To do this, portfolio managers are increasingly turning to new data sources and more sophisticated techniques to provide the edge they need to survive.

Alternative Data

Traditional data normally refers to structured data that managers have been consuming for decades. These data can easily be shown in a Microsoft Excel spreadsheet in two dimensions: Typically, time is used as one dimension, and some market or company variable is used as the other dimension. Over the past decade, new data sources have emerged to complement the traditional data sources. The four "Vs" in **Exhibit 2** can be used to describe the "bigness" of the new data. Volume refers to the exponential growth in the amount of data available. Velocity describes the speed at which data are produced and consumed. Variety describes the range of data types and sources. Finally, the fourth V, veracity, is critical because having more data is not necessarily useful unless the data are verifiable and deemed to be accurate. The data collected on smartphones illustrate the 4 Vs. Data are being created constantly on smartphones as users' apps track their various activities (volume). These data may come in text, audio, or video formats (variety). As messages and notifications are received, the user may interact with or respond to the prompts (velocity). For these data to be useful, phone manufacturers and app developers create algorithms to cleanse, store, and enrich the data (veracity). These forces are playing out across all sectors of the economy.

Asset managers have had success in collecting and incorporating these new data into their investment processes. Initial use cases include summarizing customer reviews and comments on social media and other platforms. For example, when Starbucks introduced its new rewards program in February 2016, many customers took to Twitter to protest the changes. **Exhibit 3** shows that there were significantly more "angry" and "dislike" tweets after the initial announcement and even following the rollout about one month later. Portfolio managers holding Starbucks in their portfolios could have used these trends to study the potential impact of the loyalty program changes before

Exhibit 2. The Four Vs of Alternative Data

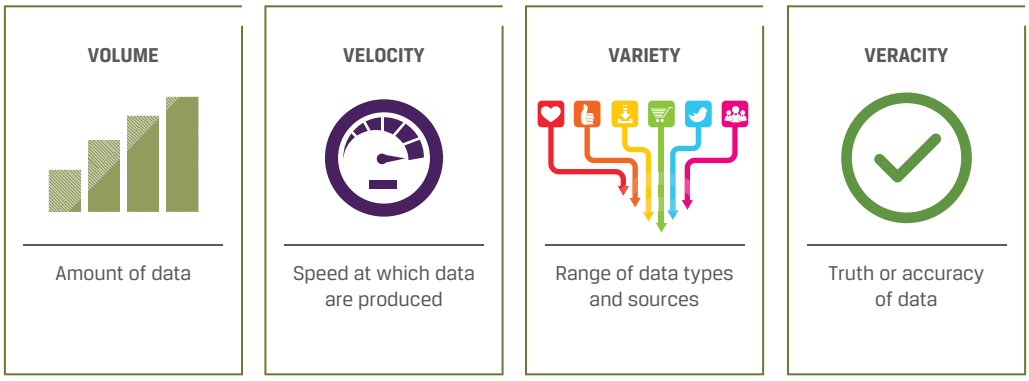
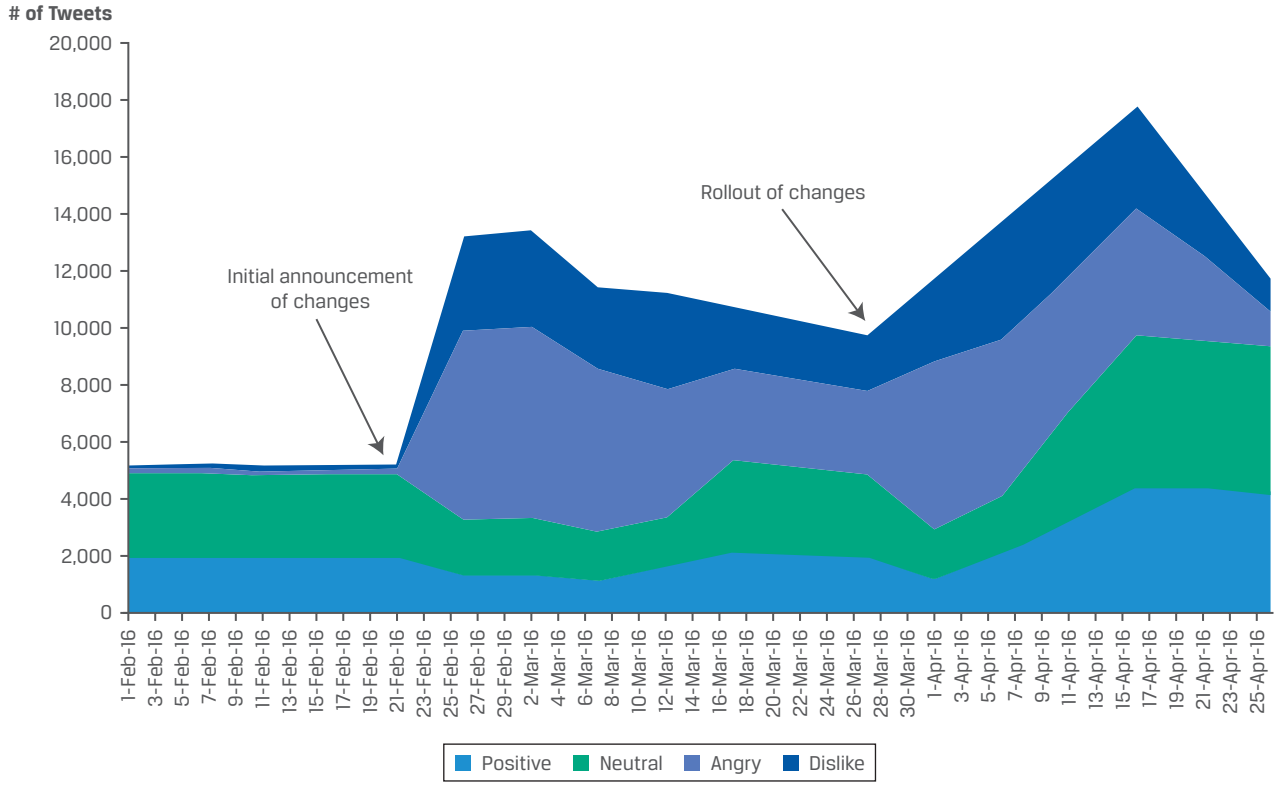


Exhibit 3. Tweets Relating to "Starbucks Rewards," 1 February 2016–25 April 2016



Source: AllianceBernstein.

the company officially reported the financial impact, likely in April 2016 to cover Q1 2016 performance. Reading and summarizing the sentiment of the tweets can give asset managers an early indication of the impact of the rewards program changes. In this case, Exhibit 3 suggests that the negative tweets had largely disappeared in April, and thus, the changes were unlikely to have a significant impact on Starbucks' financial performance.

Other applications with Twitter include measuring voting intentions during elections and monitoring product roll-outs and advertising campaigns. Beyond Twitter, product review sites and message boards may be useful for customer feedback on brands and trends. In many of these early use cases, ingesting, cleansing, and interpreting text data were key ingredients to success. Specifically, measuring sentiment and intentions across thousands

and millions of textual data requires sophisticated tools. This area is well suited for NLP.

With the evolving geopolitical and regulatory risks, companies are also looking to new data sources from governments. News and tweets can provide real-time insights into government policies and priorities, while public data sources containing infrastructure projects, shipping routes, energy consumption, enforcement actions, health data, and government spending can influence investment decisions across asset classes. Asset managers are increasingly leveraging these new datasets to enrich their understanding of the world and their impact on financial markets.

Artificial Intelligence and NLP

Artificial intelligence (AI) is the study of emulating human understanding and reaction to various situations. Machine learning (ML) is a branch of AI focused on machines learning to think by training on data. NLP is a specific form of ML that focuses on understanding and interpreting textual and spoken data. It uses linguistics, computer science, and statistics to create models that can understand text and respond to text.

NLP is a natural tool for the asset management industry because many activities of asset managers are driven by text data. Indeed, many of the alternative data trends require NLP capabilities to fully leverage their potential.

Before we discuss the applications of NLP in finance, an examination of NLP's successes and challenges in other industries can yield some insights into the evolution and adoption of these tools in our industry.

NLP Evolution and Applications

NLP research started prior to the 1950s, but the Turing test, developed by Alan Turing in 1950, was one of the first attempts to emulate human language understanding (Turing 1950, p. 457). It tests a machine's ability to "exhibit intelligent behavior" indistinguishable from humans. Initially, most of the algorithms to process language were based on predefined rules.

With the development of faster machines and the introduction of WordNet and the Penn Tree Bank in the 1980s, NLP gained prominence among researchers in computer science and in linguistics. More recently, language models incorporating neural networks, such as word2vec, allowed a vast cadre of researchers to train NLP models across different domains.

Recent NLP research relies on underlying language models to process text. A language model is a probability distribution over sequences of words. These probabilities are typically generated by training the language model on text corpora, including books, articles, news, and other forms of written text. Traditional language models include n-gram and recurrent neural network (RNN) models.

An n-gram model is a type of probabilistic model that predicts the most probable word following a sequence of *n* words. Since training corpora are typically limited, there may be many n-grams with zero probability (combinations of words that have never been seen before), especially as *n* increases. To overcome this scarcity issue, word vectors and mathematical functions can be used to capture history from text sequences and are commonly used in RNN language models.

Later implementations of RNNs use word embeddings to capture words and their positions in sequential text. These features allow RNNs to process text of varying lengths and retain longer word histories and their importance. The introduction of long short-term memory (Hochreiter and Schmidhuber 1997) in the 1990s further improved on traditional RNNs with additional capabilities to maintain information over long periods of time.

While RNNs form the backbone of many language models, there are limitations. Specifically, since text is processed sequentially, it is time consuming to train and apply these models. Moreover, when sequences are exceptionally long, RNN models may forget the contents of distant positions in sequences. Attention mechanisms (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin 2017) are commonly used to overcome this memory issue by capturing word positioning and relevance regardless of their location in sequences and allowing the neural network models to focus on the most valuable parts of the sequences. Based on attention mechanisms, transformer-based models were introduced in 2017; they vastly improved the performance of language models across various tasks. Some of the most well-known transformers include the Bidirectional Encoder Representations from Transformers or BERT (Devlin, Chang, Lee, and Toutanova 2018), Google's T5 (Raffel, Shazeer, Roberts, Lee, Narang, Matena, Zhou, Li, and Liu 2020), and OpenAI's GPT3 (Brown, Mann, Ryder, Subbiah, Kaplan, Dhariwal, Neelakantan, Shyam, Sastry, Askell, et al. 2020).

The parameters in modern NLP models are typically initialized through pretraining. This process starts with a base model, initializes the weights randomly, and trains the model from scratch on large corpora. To adjust a pretrained model for specific problems and domains, researchers fine-tune their models by further training them on the desired domain and making small adjustments to the underlying model to achieve the desired output or performance. Fine-tuning is a form of transfer learning where model parameters are adjusted for a specific task.

NLP models have been applied successfully in a variety of settings and already play important roles in our everyday lives.

Spam detection is one of the most important and practical applications of machine learning. Spam messages usually contain eye-catching words, such as "free," "win," "winner," "cash," or "prize," and tend to have words written in all capital letters or use a lot of exclamation marks. Language models can be fine-tuned to search for these common spam features to identify unwanted messages. Another popular approach for spam detection leverages supervised learning and the naive Bayes algorithm by first annotating messages as either "spam" or "not spam" and then training a model to learn from the annotations to classify new messages.

NLP models use *part-of-speech tagging* techniques that identify the nouns, verbs, adjectives, adverbs, and so on in a given a sentence. These methods enable the language models to fully understand and interpret the text.

Topic modeling using such techniques as latent Dirichlet allocation, or LDA (Blei, Ng, and Jordan 2003), have been applied extensively to extract key themes and topics in a series of sentences or texts and thus provide the ability to summarize documents quickly.

As noted earlier, data have exploded in terms of volume, velocity, and variety. Social media platforms, online forums, and company websites provide a plethora of text-based datasets that are waiting to be mined. *Sentiment analysis* can be used to analyze how different segments of the population view certain products, events, policies, and so on.

Machine language translation can be modeled as a sequence-to-sequence learning problem—that is, a sentence in the source language and another sequence returned in the translated target language. RNNs can be used to encode the meaning of the input sentence and decode the model calculations to produce the output.

Not long ago, *voice user interfaces* (VUIs) were in the realm of science fiction, but voice-enabled agents are becoming commonplace on our phones, computers, and cars. Indeed, many people may not even be aware that NLP provides the foundation for these systems. Under the hood, audio sound waves from voice are converted into language texts using ML algorithms and probabilistic models. The resulting text is then synthesized by the underlying language models to determine the meaning before formulating a response. Finally, the response text is converted back into understandable speech with ML tools.

One of the most exciting innovations in VUIs today is *conversational AI* technology. One can now carry on a conversation with a cloud-based system that incorporates well-tuned speech recognition, synthesis, and generation into one system or device. Examples include Apple's Siri, Microsoft's Cortana, Google Home, and Amazon's Alexa. The home assistant devices in this category are quite flexible. In addition to running a search or providing the weather, these devices can interface with other user-linked devices on the internet to provide more comprehensive responses. Finally, these technologies leverage cloud-based tools for speech recognition and synthesis to integrate conversational AI into many aspects of our everyday lives.

The goal behind a *question answering* (QA) system is to respond to a user's question by directly extracting information from passages or combinations of words within documents, conversations, and online searches. Almost all the state-of-the-art QA systems are built on top of pretrained language models, such as BERT (Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer, and Stoyanov 2019; Joshi, Chen, Liu, Weld, Zettlemoyer, and Levy 2020; Rajpurkar, Zhang, Lopyrev, and Liang 2016). Compared with systems that require users to scan an entire document, QA systems are more efficient because they attempt to narrow down answers quickly. Nowadays, QA systems are the foundation of chatbots, and some QA systems have extended beyond text to pictures. In some respects, QA systems can be viewed as the most generic of all the NLP solutions because the input questions can include translations, key topics, part-of-speech tags, or sentiment analysis.

NLP in Finance and Asset Management

In this section, we discuss how NLP is used in finance and asset management.

Decision-Making Framework

Given the success of NLP in solving a wide swath of problems, we now discuss its applications in finance. We start with the basic decision-making framework because it directly addresses the challenges that lead organizations toward NLP. **Exhibit 4** describes the broad actions we normally take to make decisions, regardless of whether those decisions are in finance or in everyday life. As we will see, this framework informs where NLP can be impactful.

- Gather data: Investors want to have a comprehensive view of all the data that may affect their decisions. Analysts collect financial information as well as competitor, supplier, and customer data. These data can come in many forms; alternative data sources have expanded the breadth of available data on potential investments. Whereas before, analysts would gauge real-time buying trends by visiting physical stores, they can now see more comprehensive data through geolocation or footfall data. This type of data is more

Exhibit 4. Traditional Decision-Making Framework

comprehensive because it covers many locations rather than just a hand-picked few and it tracks the trends consistently rather than for a subset of hours over a week. Ultimately, investors collect and monitor as much data as they can for their decision-making process.

- Extract insights: With reams of data at their fingertips, investors need to decide how to synthesize the data to make informed decisions. Doing so normally involves two steps: (1) determining what the data are saying and (2) deciding how to weigh each of the data points in the eventual answer.

Financial data and figures are relatively easy to interpret. Trends and other statistics can be calculated from the data to predict future performance. It is much more difficult to interpret nonfinancial data. For example, interpreting a written customer review on a website may be difficult because the source and intention of the review are both unclear. Even if we assume the review is objective and without bias, does it contain new and useful information? We normally overcome these issues by using the law of large numbers: By collecting sufficiently large sample sizes, we hope to dilute the effects of outliers and biases.

Once interpretations are extrapolated, weighing the data in the context of all the other data surrounding a potential investment is paramount. In the absence of empirical evidence, investors may use heuristics; they may assign higher weights to more recent data or data that they believe are more reliable or predictive. Asset managers tend to emphasize data that have been predictive in the past or somehow correlated with historical outcomes. These heuristics can be extremely helpful, but investors may have behavioral biases that could lead them to suboptimal results. For example, they may be biased by their connection to the CEO of a company, or they may not remember the full earnings report from last year to adequately compare the current results.

ML can overcome these issues by systematically learning from the data and monitoring the outcomes. It can be used to help weigh the data to ensure biases are mitigated and predictions are robust. Our use cases later in the article will explain these ideas in more detail.

Once the data are synthesized, investors can make their decisions using their decision-making framework.

By monitoring the outcomes (i.e., the accuracy of predictions or the impact of recommended actions), investors can then feed these new data back into the decision-making framework to improve future decisions. This iterative feedback mechanism is critical for ongoing success. ML models are trained by data and humans, so by incorporating feedback on the success or failure of the predictions, these same models can be improved. We believe this is a significant benefit of ML models. They can be trained to learn from new data and past decisions, whereas humans may be slower (or unable) to adapt to new data or incorporate failures into the decision-making framework.

NLP has a natural place in this decision-making framework because it offers a systematic approach to scan a broad set of documents and leverages ML techniques to extract insights. Reading, understanding, and synthesizing multiple documents accurately and efficiently offer clear benefits for NLP approaches.

Common NLP Tasks

We now discuss some of the common tasks relating to NLP in finance. Many of these applications were discussed in the prior section, but we will include additional comments on their uses in our industry.

Summarization

News aggregators use this technique to summarize long articles to send digestible content to inboxes. We may also wish to summarize corporate filings or presentations for quick consumption. The key challenge with this task is understanding and capturing the key relevant points of the large document. This challenge is made more difficult because different investors may weigh the content of the text differently, and as a result, the optimal summary depends on the user or usage.

Topic extraction

The key themes and topics in an article can be extracted using supervised, semi-supervised, or unsupervised methods. In the supervised approach, the model is trained to look for specific keywords related to a predefined theme, whereas the unsupervised approach attempts to infer the themes being discussed. The semi-supervised approach

is a hybrid approach in which seed words can be used as starting points to identify themes. One example is in the context of central bank statements. Common themes debated during central bank meetings in the United States include inflation and growth. We may seed the topic modeling process with these words and then apply various language techniques to find new words and phrases resembling these themes to derive the final topics from the documents.

Search/information retrieval

We may be interested in looking for specific terms or references from the text. Examples may include competitor or product references in a document. An analyst for Apple may be interested in finding all references to the iPhone in corporate documents or news articles.

Question answering

Similar to the search task, we are interested in finding information in a document. However, the QA task is focused on answering specific questions rather than returning references to those questions. In the previous example regarding the Apple analyst, he may be interested in the actual iPhone units sold rather than simply references to the phone. Another example may be extracting the interest rate or the covenants in a loan document. A specific application of this question answering task is the chat box, where specific and relevant responses are needed to reply to questions.

Sentiment analysis

Should the text be viewed positively or negatively? This will again depend on the user and application. For example, certain words, such as "debit," "liability," and "resolve," may have different meanings depending on the context. The first two terms are normally viewed as negative words in common parlance but tend to be more neutral in a financial setting since they are common terms in financial statements. "Resolve" may be viewed as positive in most settings, but "did not resolve" should be viewed as negative.

Named entity recognition

Extracting entity names from text is a common but important task. Entities include countries, organizations, companies, individuals, places, and products. By identifying the entities in a document, investors can link the article to other information on the entities and thereby create a comprehensive view before making decisions. Client interactions are no different: Determining client-relevant text is critical to a complete understanding of a client. While this type of data is critical, the problem is difficult; for example, separating articles regarding the technology company Apple and the fruit apple is not trivial.

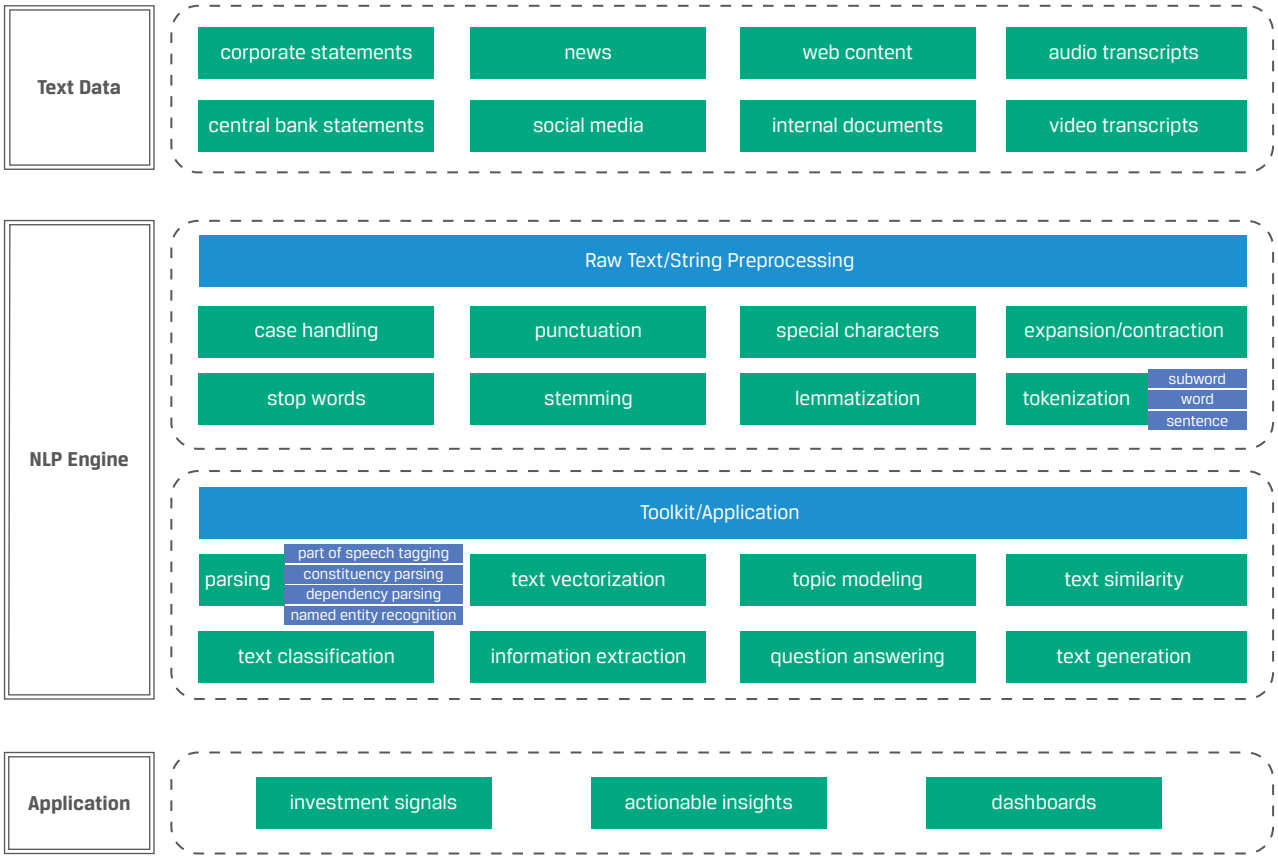
Typical NLP Pipeline

With these applications in mind, we discuss the infrastructure required to incorporate NLP at scale. The NLP pipeline depicted in **Exhibit 5** presents a high-level overview of the key components in NLP analysis that mainly handle text inputs. Note that the original data need to be converted to machine readable text. For example, speech to text is required for audio and video sources. The original data can come from a variety of external and internal sources, such as corporate statements, central bank statements, news, social media, web content, internal documents, and video/audio meeting transcriptions. A data lake is typically needed to handle large amounts of input data. Depending on the size, structure, and usage of the data, they can be stored in different types of databases. Data with clear structure or schema can typically be stored using SQL relational databases. Unstructured data can be stored in nonrelational databases, such as MongoDB. Today, graph databases are becoming more popular in handling exceptionally large sets of structured, semistructured, or unstructured data.

The NLP engine shown in Exhibit 5 depicts the tools for text processing and is composed of two components. The first part focuses on processing raw text. Different use cases require different preprocessing procedures. For example, "case handling" refers to the conversion of all characters to uppercase/lowercase. This process is nuanced because "us" can refer to the United States if written in uppercase or the pronoun "us" if lowercase is used. In addition, while the removal of irregular punctuation and characters from formal documents, such as corporate filings, may be warranted, some characters may carry valuable information in other domains, such as emojis in social media data. Expansion/contraction refers to the process of expanding contractions into separate words, such as replacing "isn't" with "is not." This procedure will affect the final list of tokenized words (the result of splitting text into smaller units). Stop words, such as "the," "a," and "is," are not informative and are typically removed to reduce the size of the vocabulary. Depending on the application, stemming or lemmatization may be applied to reduce words to their root or stem forms. In the tokenization procedure, text in a document can be broken down into sentences, and these sentences can be further split into words or subwords. In some situations, subwords may be more robust when dealing with rare words.

The second part of the NLP engine contains tools to process and transform the cleansed text into usable information that can be consumed by the end applications or users. For example, we can analyze a sentence by splitting it into its parts and describing the syntactic roles of the various pieces. Named entity recognition is a commonly used parsing technique to identify persons, locations, and organizations from the text. Text vectorization—using

Exhibit 5. NLP Pipeline



vectors to represent text—is commonly used for feature engineering. Popular techniques for text vectorization include one-hot encoding, term frequency–inverse document frequency (TF–IDF), word embeddings using word2vec (Mikolov, Chen, Corrado, and Dean 2013) and GloVe (Pennington, Socher, and Manning 2014), and word/sentence embeddings with deep learning language models. As an example, representing documents through bag-of-words approaches using one-hot encoding allows us to compute statistics describing the words in the document. These statistics can be enhanced to carry more information by assigning weights to the various words using such methods as TF–IDF. Word embeddings using word2vec or GloVe allow us to represent individual words using numerical vectors. Deep learning language models, such as BERT and GPT-3, can represent words and sentences using numerical vectors containing rich contextual information, such as positioning within sentences. These models are very efficient and powerful and have a wide range of applications.

Asset managers should modularize their tools to leverage these tools on diverse types of documents. For example, text summarization tools should be abstracted so they can be used for news articles, corporate filings, and internal emails. By modularizing the code, the various NLP techniques can be applied broadly across different types of documents. Depending on the size of the data and the specific algorithms used, different infrastructure may be needed. For example, to create sentence embeddings from many documents, a single GPU (graphics processing unit) processor can be substantially faster than parallelizing multiple CPUs (central processing units). For even larger corpora, such as global news articles, large clusters of machines may be needed to create deep learning models.

The output from the NLP engine can be consumed in different ways. For example, the NLP signals can be used directly to prompt an investment action or as part of a broader strategy. There may also be value in showing the results in a dashboard where users can interact with the original documents and the output. This transparency gives the users confidence in the signals because they can easily review the results on their own. In addition, asset managers may leverage these dashboards to solicit feedback from their users to improve their algorithms. One idea may be to give users the ability to highlight certain text and annotate the sentiment of the text to further fine-tune their models.

NLP Applications in Asset Management

In this section, we examine several applications of NLP in the asset management industry.

Deciphering Sentiment in Social Media Posts

Let us start with a simple example of applying NLP to social media posts to track sentiment. When the iPhone X was launched in late 2017, there were questions about its adoption and success. Investors and analysts turned to social media posts and reviews to gauge the sentiment of the iPhone X, as well as its competitors, during the rollout.

Investors were interested in understanding the features consumers valued in smartphones and whether the iPhone X's new features (Face ID, lack of a home button, etc.) were resonating with consumers.

One common approach to this question was to scrape smartphone reviews among various websites to assess consumer feedback. A simple word cloud, such as the one shown in **Exhibit 6**, is a quick way to uncover the key topics and themes.

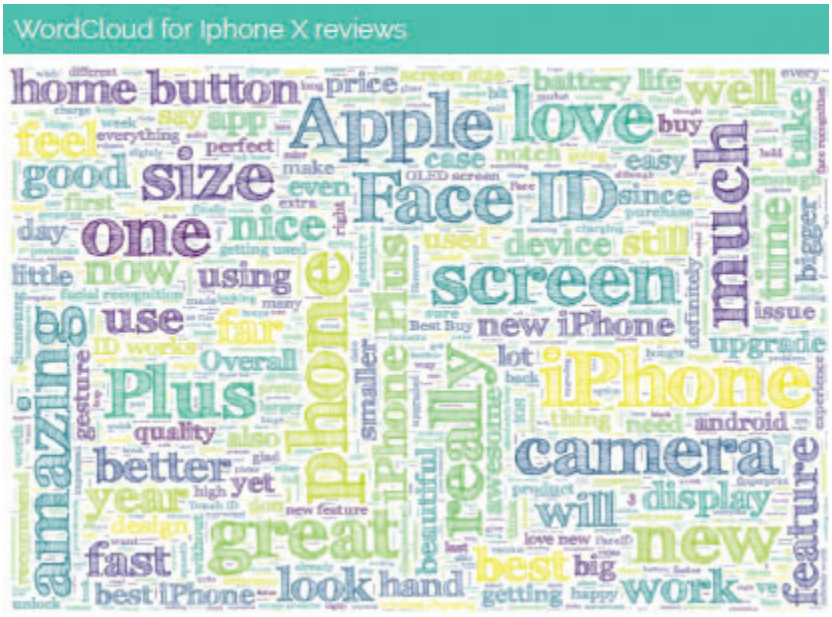
The larger words represent the most frequently used words in the reviews. While some words are meaningless ("much" and "use") or expected ("Apple"), it may have been reassuring to see that "Face ID" was being highlighted and such words as "amazing" and "love" were being used to describe the new phone.

Investors can also analyze the sentiment across the reviews to gauge consumer feedback on the various smartphone models. VADER (Valence Aware Dictionary and sEntiment Reasoner) is well suited for this purpose given its ability to measure polarity and intensity of social media posts (Hutto and Gilbert 2014). **Exhibit 7** shows a snapshot of the reviews soon after the iPhone X was introduced. While most of the consumers were "neutral" on the new phone, it seemed that the iPhone X was lagging its predecessors in terms of positive reviews.

Investors may have also wanted to determine the features that were most prevalent in reviews to assess the appeal of the iPhone X's new features since those enhancements were likely going to drive adoption. This analysis typically requires the determination of the key features differentiating the phones (system, screen, battery, price, etc.) and the creation of a list of keywords used to describe those features. For example, such terms as "iOS" and "Android" are associated with the "system" feature. **Exhibit 8** suggests that the new features in the iPhone X were leading to many positive reviews.

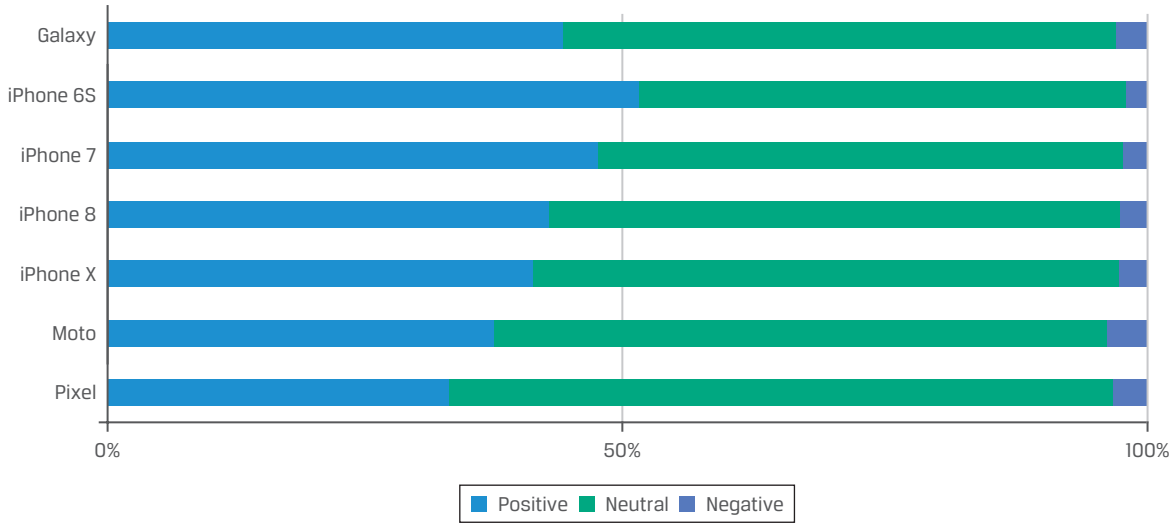
This simple example illustrates the basics of NLP analysis. Once the raw data from social media posts and customer reviews are ingested, simple tools can be used to analyze the data to provide insights into investment controversies. In late 2017, investors were assessing the launch of the iPhone X, and by monitoring and

Exhibit 6. Word Cloud for iPhone X Reviews



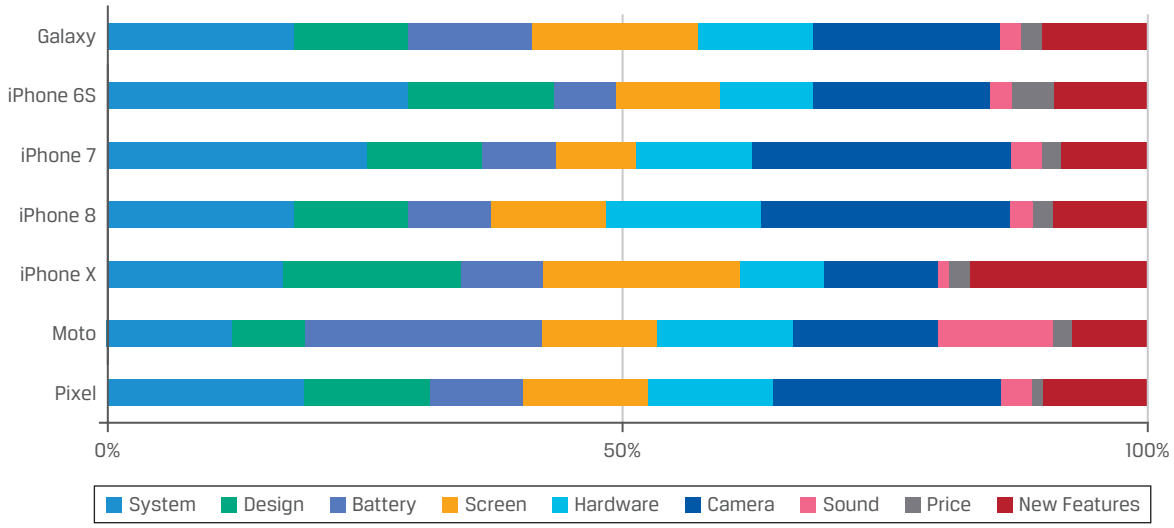
Source: AllianceBernstein.

Exhibit 7. Consumer Sentiment by Phone



Source: AllianceBernstein.

Exhibit 8. Percentage of Feature Mentions in Positive Reviews



Source: AllianceBernstein.

aggregating the social media posts, they were likely able to discern the positive adoption of the new features and thus conclude that the new phone was likely to achieve the success of the prior models.

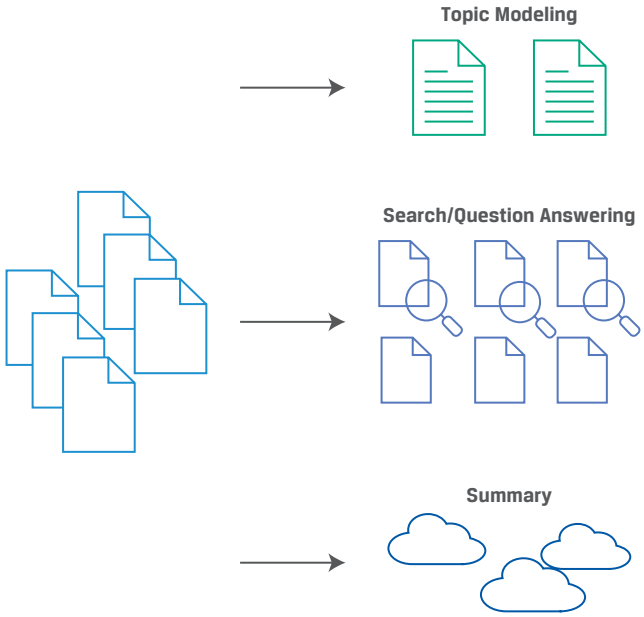
Extracting Themes to Highlight Important Topics and Trends

Our industry deals with vast amounts of documents. Often, analysts are tasked with synthesizing and summarizing enormous documents or extracting important sections or figures from these documents. **Exhibit 9** highlights some of the key tasks when dealing with large documents.

Topic modeling can streamline the analysis of large documents by identifying and extracting the key topics or themes in the data. The evolution of these topics from the corpora may also provide insight into the importance of the themes over time.

In the following example, the BERTopic package (Grootendorst 2022) is used to generate topic representations by converting each document to its embedded

Exhibit 9. Synthesizing Large Documents

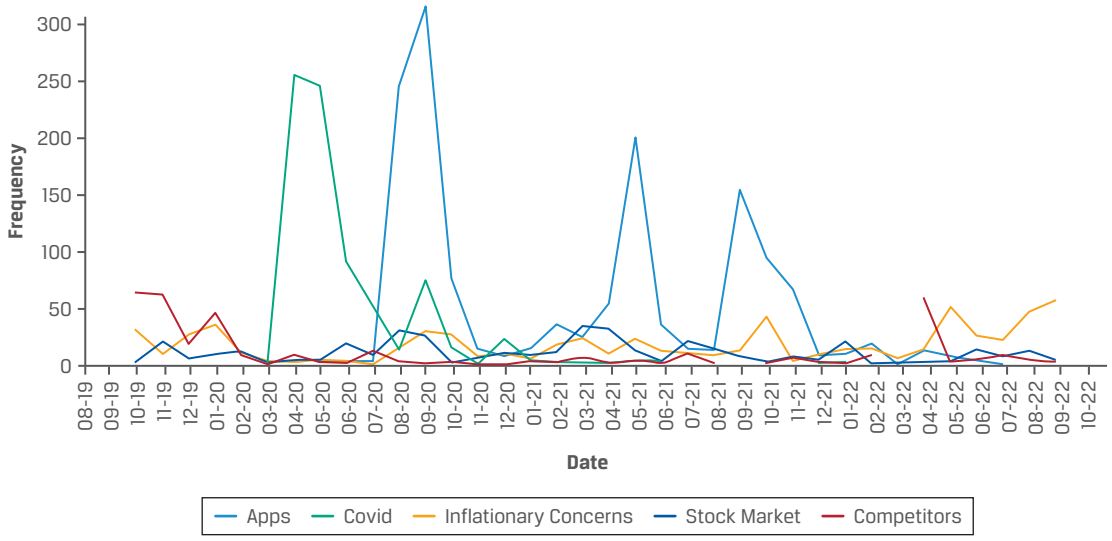


representation, clustering the documents, and then extracting keywords to represent the derived themes.

We applied BERTopic to about 260,000 Apple headlines from October 2019 to September 2022 to extract the top topics from the text. Since the topics were generated programmatically, we inferred the true themes for each of the topics from the generated keywords and used them as labels in **Exhibit 10**. For example, we created a theme called "Stock Market" because the important keywords for the topic are "sp," "500," "nasdaq," and "dow." Similarly, the "Covid" theme has references to "infection," "contact," and "tracing." Deciphering intuitive and interpretable themes from the BERTopic output is a crucial step when using topic modeling tools.

The topics we inferred from the Apple news articles ("Apps," "Covid," "Inflation Concerns," "Stock Market," and

Exhibit 10. Top Five Topics from Apple News Articles, October 2019–September 2022



Source: AllianceBernstein.

"Competitors") are intuitive and distinct. Note that "Covid" was unsurprisingly the key topic in the first half of 2020, but its importance has waned since then. Instead, inflationary concerns are dominating more recent Apple news headlines. **Exhibit 11** provides the sample code to extract key themes from text.

Topic modeling is particularly powerful in uncovering the key themes or ideas across a set of documents. It can also be used to explain the evolution of topics over time to understand their importance in a specific area. In the Apple example, investors can extract the key themes and the associated keywords from market news to drive their research and perspective on the company. In another example, we applied topic modeling on central bank statements to uncover the emphasis of the Federal Open Market Committee (FOMC) on its two goals—price stability and sustained economic growth. By parsing central bank statements, we could assess the FOMC's trade-off between these two goals to gain insights into their policies at upcoming meetings.

Exhibit 11. Sample Code: Extracting Key Themes

Load in time series data - Apple headlines

```
import pandas as pd
apple_news = pd.read_csv("./apple_news.csv").dropna()
apple_news = apple_news.drop_duplicates( keep='last')
apple_news['date']= pd.to_datetime(apple_news['date'])
apple = apple_news.groupby(['date'])['text'].apply(lambda x: ' '.join(x)).to_frame().reset_index()
timestamps = apple.date.to_list()
apple_news_list = apple.text.to_list()
apple_news.head(3)
```

	date	text
0	2019-09-22 11:00:00	Apple (AAPL) Valuation Rose While Independent ...
1	2019-09-22 11:00:00	Do Directors Own Apple Inc. (NASDAQ:AAPL) Shares?
2	2019-09-22 11:45:00	Global tax authorities discuss targeting multi...

Calculate headline embeddings and create clusters over time

```
from bertopic import BERTopic
# Initialize the model
topic_model = BERTopic(verbose=True)
# calculate headline embeddings
topics, probs = topic_model.fit_transform(apple_news_list)
# create clusters over time
topics_over_time = topic_model.topics_over_time(apple_news_list, topics, timestamps, nr_bins=36)
top_topics = topics_over_time[topics_over_time['Topic'].isin(range(5))]
top_topics = top_topics.set_index('Timestamp')
top_topics_pivot = top_topics.pivot_table(index='Timestamp',columns='Topic',values='Frequency',aggfunc='sum')
```

Visualize the top 5 topics over time

```
import matplotlib.pyplot as plt
import matplotlib.dates as mdates
fig = plt.figure(figsize=(25,10))
ax = plt.axes()
# re-name topic names
plt.plot(top_topics_pivot[0], label='Apps', color='purple',linewidth=3.0)
plt.plot(top_topics_pivot[1], label='Covid', color='black',linewidth=3.0)
plt.plot(top_topics_pivot[2], label='Inflationary Concerns', color='blue',linewidth=3.0)
plt.plot(top_topics_pivot[3], label='Stock Market', color='green',linewidth=3.0)
plt.plot(top_topics_pivot[4], label='Competitors', color='red',linewidth=3.0)
# change background color and date format
ax.set_facecolor("white")
ax.xaxis.set_major_locator(mdates.MonthLocator())
ax.xaxis.set_major_formatter(mdates.DateFormatter('%m-%y'))
# add axes labels and a title
plt.title('Top 5 Topics Over Time\n', fontsize=18)
plt.xlabel('\n Date', fontsize=16)
plt.ylabel('\n Frequency', fontsize=16)
# display plot with legend
plt.legend(title='Topic_Name')
plt.show()
```

Searching for Key Themes and Question Answering

While topic modeling can provide a good overview of a set of documents, investors may be interested in extracting relevant sections or targeted data points from specific documents. We provide an example of these tasks using environmental, social, and governance (ESG) documents. Because ESG-oriented strategies have become more mainstream, asset managers are looking for ways to assess ESG-related activities in their investment companies and to monitor their progress toward their goals. Corporate and social responsibility (CSR) reports are used by companies to communicate their ESG efforts and their impact on the environment and community. These reports describe the company's relations with its full range of stakeholders: employees, customers, communities, suppliers, governments, and shareholders. Though corporations are currently not mandated to publish CSR reports annually, more than 90% of the companies in the S&P 500 Index did so for 2019.

In the following example shown in **Exhibit 12**, we parse sentences from a CSR report in the automotive industry and leverage a semantic search model that uses predefined keywords to rank and select parsed sentences. This example searches for sentences related to "GHG [greenhouse gas] Emissions." We embed this phrase into a vector and use it to compare against the embedded representation of the document text. Our example uses only one comparison sentence ("text" variable in the sample code below) for simplicity, but we can apply the same process for multiple candidate sentences, sort similarity scores across all of them, and then select the most similar ones from the process.

Exhibit 12. Sample Code: Searching for ESG Themes in a CSR Report

Load packages and the USE model

```
import numpy as np
import tensorflow as tf
import tensorflow_hub as hub
module_url = "https://tfhub.dev/google/universal-sentence-encoder/4"
use_model = hub.load(module_url)
```

Specify the keyword and text, and calculate embeddings separately

```
keyword = "GHG Emissions"
keyword_vec = use_model([keyword])[0]

text =['''Energy indirect (Scope 2) GHG emissions Baseline year 2010,
which was the first full year of operation as the new
General Motors Company and includes all facilities under GM operational control.
Calculation includes CO2, CH4 and N20.
Reporting is based on GHG Protocol, and the source of emission factors is regulatory or IPCC.
2020 GHG emissions are as follows:Gross location based indirect emissions: 3,087,816 Metric tons CO2e
Gross market based indirect emissions: 2,599,822 Metric tons CO2e''']
sentence_embeddings = use_model(text)
```

Calculate the cosine similarity

```
def calculate_cosine_similarity(u, v):
    return np.dot(u, v) / (np.linalg.norm(u) * np.linalg.norm(v))
score = calculate_cosine_similarity(keyword_vec, sentence_embeddings[0])
```

Print out results

```
print("Query = ", query,end='\n\n')
print("Sentence = ", text[0],end='\n\n')
print("Similarity Score = ", round(score,2))
```

```
Query =  GHG Emissions

Sentence =  Energy indirect (Scope 2) GHG emissions Baseline year 2010, which was the first full year of operation as the ne
w
General Motors Company and includes all facilities under GM operational control. Calculation includes CO2, CH4 and N20.
Reporting is based on GHG Protocol, and the source of emission factors is regulatory or IPCC. 2020 GHG emissions are as foll
ows:
Gross location based indirect emissions: 3,087,816 Metric tons CO2e Gross market based indirect emissions: 2,599,822 Metric
tons CO2e

Similarity Score =  0.46
```

Theme searches can direct researchers to relevant sections of the text quickly, thus simplifying the process to extract insights from documents. One caveat is that semantic searches initially start as unsupervised learning processes and may need to be enhanced by creating classification models based on labeled feedback. These resulting supervised learning models ensure important sections are not missed during the search process. Indeed, creating a comprehensive list of keywords and fine-tuning the model on annotated search results are common methods to minimize false negatives (incorrectly labeling a section as not important).

While finding relevant and related text improves both the efficiency and effectiveness of researchers, more direct approaches to narrow down the answers to queries may be even more impactful. Question answering is designed for exactly this purpose. The goal of QA is to build systems that automatically extract answers from a given corpus for questions posed by humans in a natural language. In the following example in **Exhibit 13**, we feed the question "What is the goal by 2035?" and a representative passage from a CSR report into the RoBERTa model (Liu et al. 2019), an optimized model leveraging BERT. The model can extract the exact answer from the passage—"source 100% renewable energy globally"—thus directly answering the original question. In real-life examples, passages and documents are much longer but the same approach can be used to find the best answer to the question.

These techniques have broad applications across the asset management industry; research analysts, risk managers,

Exhibit 13. Sample Code: Question Answering for ESG Metrics

Load packages and the RoBERTa models

```
import torch
from transformers import AutoTokenizer, AutoModelForQuestionAnswering
tokenizer = AutoTokenizer.from_pretrained("vanadhi/roberta-base-fiqa-flm-sq-flit")
model = AutoModelForQuestionAnswering.from_pretrained("vanadhi/roberta-base-fiqa-flm-sq-flit")
```

Specify the question and text, and calculate embedding inputs using the tokenizer

```
question ="What is the goal by 2035?"

text = '''We're committed to achieving this vision in a timeframe that aligns with climate science.
That's why GM has announced plans to become carbon neutral in our global products and operations by 2040.
Making progress toward these goals will address the most significant sources of
carbon emissions that we may be able to impact, including vehicle emissions, which currently represent 75% of the
emissions we are trying to reduce, and our manufacturing operations, which are responsible for 2%. To reach carbon
neutrality in our operations, we have a goal to source 100% renewable energy globally by 2035, five years earlier than
our previous commitment made in 2020 and 15 years sooner than our original target.'''

inputs = tokenizer(question, text, return_tensors='pt')
```

Feed input embeddings into the model, and extract an answer from the text

```
outputs = model(**inputs)
start_scores = outputs.start_logits
end_scores = outputs.end_logits
answer_start = torch.argmax(start_scores)
answer_end = torch.argmax(end_scores) + 1
answer = tokenizer.convert_tokens_to_string(
    tokenizer.convert_ids_to_tokens(inputs["input_ids"][0][answer_start:answer_end]))
```

Print out results

```
print('Question: '+ question, end='\n\n')
print('Context: '+ text, end='\n\n')
print('Answer: '+answer)
```

```
Question: What is the goal by 2035?

Context: We're committed to achieving this vision in a timeframe that aligns with climate science.
That's why GM has announced plans to become carbon neutral in our global products and operations by 2040.
Making progress toward these goals will address the most significant sources of
carbon emissions that we may be able to impact, including vehicle emissions, which currently represent 75% of the
emissions we are trying to reduce, and our manufacturing operations, which are responsible for 2%. To reach carbon
neutrality in our operations, we have a goal to source 100% renewable energy globally by 2035, five years earlier than
our previous commitment made in 2020 and 15 years sooner than our original target.

Answer:  source 100% renewable energy globally
```

compliance officers, and operations staff are constantly scouring documents for key figures and specific items within documents. Examples include financial statement analysis, ESG monitoring, regulatory reporting, and fund prospectus reviews. While these tools can be extremely powerful, they need careful calibration and fine-tuning before they can be used widely across the financial services industry. In our experience, creating a step to extract relevant sections *before* the QA process, as in the prior use case, is essential to success. This step ensures that the appropriate sections of the document are being searched for the answers.

Uncovering Risks in Corporate Filings

We now delve further into the actual text and language used in documents to uncover insights for investment research. We scan corporate filings for potential risks using text mining techniques. Since the Securities and Exchange Commission (SEC) requires publicly traded companies to file reports disclosing their financial condition, investors can parse these reports for significant disclosures, changes, and trends.

In an influential paper, Cohen, Malloy, and Nguyen (2019) found that significant changes in sequential 10-Ks convey negative information on future firm performance. Since 10-K filings are typically long and complicated, leveraging NLP techniques to extract information systematically can greatly improve the comparison of year-over-year (YOY) changes. Cohen et al. used several bag-of-words approaches to measure document changes, including the cosine similarity between vectors representing the documents. The following example in **Exhibit 14** leverages doc2vec (a generalized version of word2vec that represents whole documents as vectors) to model the changes of the management discussion and analysis (MD&A) section in 10-K forms.

Corporate filings, including 10-Ks, can be scraped from the SEC's websites. Text from the MD&A section is extracted and doc2vec is used to represent the text in each of the filings. Specifically, the words in the MD&A sections are represented by numerical vectors using the doc2vec algorithm. To gauge YOY changes, we compute the cosine similarity of the two vectors representing the sequential filings. High cosine similarity suggests the text from the two filings is largely the same, whereas low cosine similarity suggests substantial differences in the underlying reports.

A typical long–short backtest can be used to determine the predictive efficacy of this NLP feature. **Exhibit 15** shows the performance of a monthly-rebalanced long–short strategy, with the long side representing the companies with the most similar YOY filings and the short side containing the companies with the most dissimilar filings. The backtest shows compelling results throughout the study period, with the strongest results in more recent years. Our results suggest that simple calculations using vectorized representations of documents can uncover risks and opportunities from corporate filings. The NLP process in this example "reads" the documents and looks for differences in the text, emulating the tasks commonly performed by research analysts.

Broadening Insights on Earnings Call Transcripts

Earnings calls provide forums for companies to convey important financial and business information to the investment community and the public. Human analysts scour individual calls for insights into company operations, but doing so systematically across a wide swath of companies can be time consuming and error prone. NLP tools can be leveraged efficiently and effectively to address these issues.

Earnings calls typically consist of a presentation section and a question-and-answer (Q&A) section. Company executives are the sole participants in the first section, whereas both corporate executives and analysts from the investment community interact during the Q&A section. Investment signals can be mined on the different sections and the different types of speakers—namely, CEOs, other executives, and analysts—to study potential differences among them.

A variety of NLP approaches can be used to generate investment signals from earnings call transcripts. Bag-of-words approaches using predefined dictionaries and context-driven language models are common techniques. We describe three categories of features in our analysis of the transcripts—document attributes, readability scores, and sentiment scores.

Document attributes refer to features derived from the characteristics of the call. Examples include the number of words, sentences, questions, and analyst participants in a call.

Readability scores use a variety of methods to assess the difficulty of the text and document. These metrics tend to focus on two areas: the use of difficult-to-understand words and the length of sentences. Easily understood messages (texts with low readability scores) may be quickly incorporated into market prices and will therefore have a negligible impact on potential mispricing. Complex messages (texts with high readability scores) may be used by company executives to obfuscate bad news or less-than-stellar results.

Sentiment scores can be derived from the underlying text using different formulations. The most basic method to assess sentiment is to count the number of positive and negative words based on a specific dictionary, such as Harvard IV-4, VADER (Hutto and Gilbert 2014), and Loughran–McDonald (Loughran and McDonald 2011). This approach, commonly called bag of words or dictionary

Exhibit 14. Sample Code: Determining Changes in Corporate Filings

Load packages

```
import pandas as pd
import numpy as np
from nltk.tokenize import word_tokenize
from gensim.models.doc2vec import Doc2Vec, TaggedDocument
```

Prepare data

```
df = pd.read_csv('df_AAPL_10K_MDA.csv')
train = list(df[~df['heading'].str.contains('highlights')]['sectionText'].dropna())
test = list(df[df['heading'].str.contains('highlights')]['sectionText'].values)
print("There are {} pieces of text used for model training and {} pieces of text used in test"\
      .format(len(train), len(test)))
```

```
There are 248 pieces of text used for model training and 9 pieces of text used in test
```

Train a doc2vec model based on training data

```
# Tokenize and tag each document
train_tokenized = [word_tokenize(doc.lower()) for doc in train]
train_tagged = [TaggedDocument(d, [i]) for i, d in enumerate(train_tokenized)]

# Train doc2vec model
'''
vector_size = Dimensionality of the feature vectors.
window = The maximum distance between the current and predicted word within a sentence.
min_count = Ignores all words with total frequency lower than this.
alpha = The initial learning rate.
'''
model = Doc2Vec(train_tagged, vector_size = 32, window = 3, min_count = 1, epochs = 100)
```

Get the results on the test data

```
# tokenzize test docs
test_tokenized = [word_tokenize(doc.lower()) for doc in test]

# get the vector representation for the test docs
test_vectors = [model.infer_vector(doc) for doc in test_tokenized]
n = len(test_vectors)
```

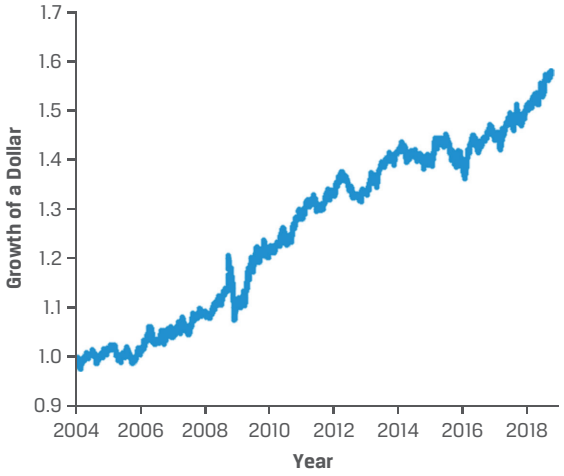
Calculate the cosine similarity and print out results

```
def calculate_cosine_similarity(u, v):
    return np.dot(u, v) / (np.linalg.norm(u) * np.linalg.norm(v))

YoY_sim = []
for i in range(len(test_vectors)-1):
        YoY_sim.append(calculate_cosine_similarity(test_vectors[i], test_vectors[i+1]))

print(pd.Series(YoY_sim, index=[str(year-1)+'-'+str(year) for year in range(2014, 2022)]).round(2))
```

```
2013-2014    0.93
2014-2015    0.94
2015-2016    0.88
2016-2017    0.94
2017-2018    0.95
2018-2019    0.79
2019-2020    0.59
2020-2021    0.90
```

based, is intuitive and interpretable, but it has limitations. For example, it cannot handle negation or words that may have different meanings in different settings. Context-driven language models can overcome the issues of dictionary-based approaches. With the development of advanced algorithms and improvements in computational power, transformer-based models, such as BERT, have proven to be effective in encoding and decoding the semantic and syntactic information of natural languages. The sample code in **Exhibit 16** uses FinBERT (Huang, Wang, and Yang, forthcoming), a BERT-based model pretrained on financial text, to score sentiment on a series of input sentences.

Exhibit 15. Performance of Similarity Score on US Large-Cap Companies, 2004–2019



Source: AllianceBernstein.

We assess the ability of our features to differentiate between outperforming and underperforming stocks using a monthly rebalanced strategy. At the beginning of each month, we form portfolios based on the available features as of that date and track the difference in subsequent returns between the top and bottom quintiles over the following month. **Exhibit 17** shows a histogram of information ratios (IRs) for various features on US large-cap companies over the period 2010–2021. Approximately 20% of the IRs in our research are greater than 0.5 for the US companies, suggesting there is promise in the features. We found similarly encouraging results for features created in other stock and bond universes.

We now analyze the differences between dictionary-based and context-driven approaches to derive sentiment. Conceptually, the context around words should be an important driver when assessing sentiment. While "grow" may typically be viewed as positive, "competitors growing" or "economic headwinds growing" should be scored negatively. **Exhibit 18** shows the dollar growth of the strategies based on representative sentiment scores generated through these two approaches. The context-driven approach based on BERT has performed better recently, suggesting that it has been able to better discern sentiment in financial text. Additionally, there has been evidence suggesting company executives are adapting to the rise of machines listening to their comments by changing the words they use in their communications. In other words, company executives may be choosing their words carefully since they know NLP techniques are being used to analyze their comments. Thus, dictionary-based approaches may not be as useful going forward, and context-driven approaches may be more robust in overcoming these behavioral changes.

These high-level results suggest that the generated NLP signals from earnings call transcripts may be useful for portfolio managers. Applications include the usage of these signals in systematic strategies or to complement fundamental processes. Additionally, similar techniques can be applied to other types of corporate filings and statements to extract insights from those documents. These tools give investors the ability to analyze large amounts of documents methodically and potentially save them from doing the work manually.

Deepening Client Insights to Prioritize Sales Efforts

With continued economic and market uncertainty, sales teams need to quickly digest and synthesize client information and updates to assess their needs. We can leverage NLP to help these sales teams deepen client insights and prioritize sales efforts using publicly available data.

The data sources include client presentations, quarterly/annual reports, meeting minutes, announcements, and news. It is time consuming for any sales team to monitor all the media outlets for information across a broad set of prospects and clients. NLP techniques can be leveraged to collect, process, and synthesize the data and alert the sales team with timely and actionable prompts.

Exhibit 19 illustrates this process. First, public data can be obtained through web scraping. For example, data scraping pipelines can be built to detect document changes on various websites. We can also monitor major news outlets to track mentions of specific keywords, such as names of organizations, themes, and topics. The scraped data are populated into a database before being fed into the NLP engine. Based on the specific use case, the NLP engine (more details are shown in Exhibit 5) aims to further prepare the data and applies the applicable algorithms to extract the relevant information. As a final step, notifications and prompts are sent to the sales team for further action.

In summary, using NLP to effectively process and synthesize information can inform and improve sales outreach by surfacing timely alerts and relevant intelligence from publicly available data sources.

Identifying Entities within Documents

Across many of the use cases discussed above, entity names and identifiers need to be extracted from the text to tie the documents back to specific people or organizations.

Exhibit 16. Sample Code: Extracting Sentiment from Text

Load packages and the finBERT models

```
from transformers import BertTokenizer, BertForSequenceClassification
from transformers import pipeline
import pandas as pd

finbert = BertForSequenceClassification.from_pretrained('yiyanghkust/finbert-tone',num_labels=3)
tokenizer = BertTokenizer.from_pretrained('yiyanghkust/finbert-tone')

nlp = pipeline("sentiment-analysis", model=finbert, tokenizer=tokenizer)
```

Predict the sentiment of sentences

```
sentences = ["there is a shortage of capital, and we need extra financing.",
             "growth is strong and we have plenty of liquidity.",
             "there are doubts about our finances.",
             "profits are flat."]
results = nlp(sentences)
```

Check the results

```
%%capture
[results[i].update({'sentence': sentences[i]}) for i in range(len(sentences))]
output = pd.DataFrame(results)[['sentence', 'score', 'label']]
output['score'] = output['score'].round(3)
```

```
output
```

	sentence	score	label
0	there is a shortage of capital, and we need extra financing.	0.995	Negative
1	growth is strong and we have plenty of liquidity.	1.000	Positive
2	there are doubts about our finances.	1.000	Negative
3	profits are flat.	0.994	Neutral

Named entity recognition refers to the task of identifying the entities (person, place, country, company, financial instrument, etc.) within documents. **Exhibit 20** provides a simple example using spaCy to identify the various entities in a news title.

Successfully Leveraging NLP

While our use cases demonstrate that significant advances have been made, NLP is still relatively new in finance. We discuss the key technical and business challenges next.

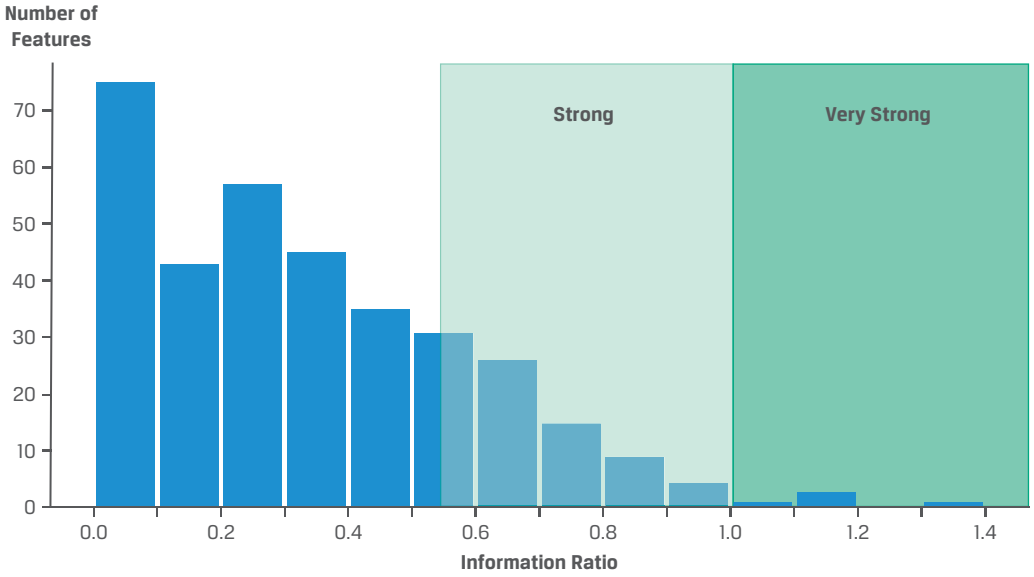
Overcoming Technical Challenges

Broadly, technical challenges, such as breaking down sentences and tagging the parts of speech within sentences, are common across all NLP applications. Linguists and computer scientists are researching and creating new insights to improve on these capabilities, especially across different languages.

Context-specific tools for the financial services industry are being developed. Dictionaries are extremely useful and improve language understanding, especially for specialized domains. One complexity lies in robust dictionaries across different languages. We found that a simple translation of the Loughran–McDonald dictionary to Chinese is not robust because some words do not carry the same meanings in Chinese and there are additional Chinese words that are more meaningful for Chinese investors.

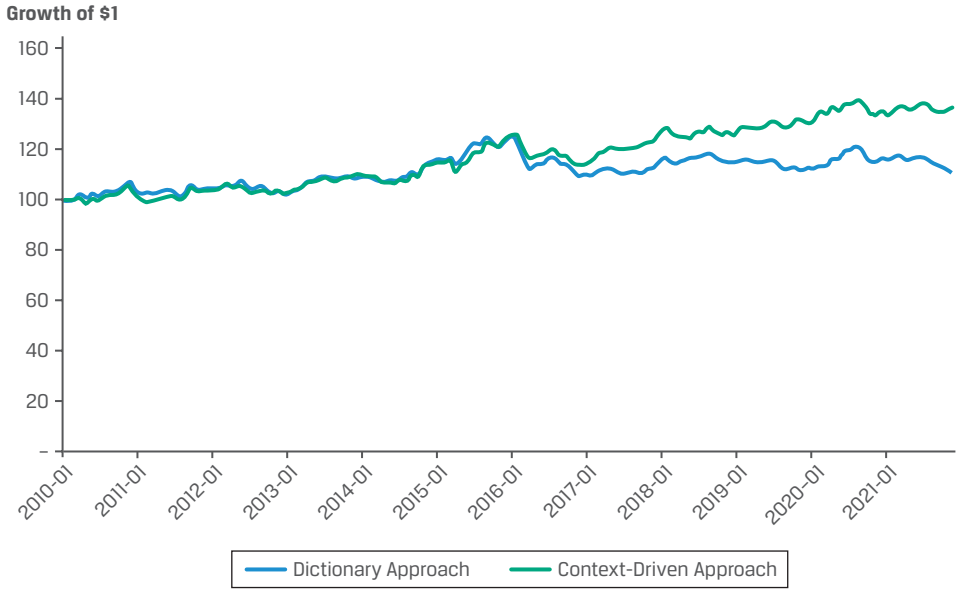
Even context-specific language models, such as BERT, will need to learn from new financial domains. While FinBERT-type models have been trained on financial text, further training and fine-tuning is likely required to improve

Exhibit 17. Number of NLP Features across Different Information Ratio Levels



Source: AllianceBernstein.

Exhibit 18. Performance of Context-Driven vs. Dictionary-Based Approaches, 2010–2021



Source: AllianceBernstein.

performance in specialized areas, such as securitized assets, regulatory filings, fund disclosures, and economic forecasts. In securitized assets, for example, such common English words as "pool" and "tranche" have vastly different meanings, and language models will need to be trained with these new words before machines can systematically understand and synthesize text as humans do.

Named entity recognition (NER) is an important task in finance. Even though we provided a simple example among our use cases to apply spaCy's NER model on written text, the model is not perfect. For example, spaCy's model incorrectly tags companies with names containing common English words. For example, numerous articles containing the words "state" or "street" may be classified under

Exhibit 19. Overview of Process to Deepen Client Insights

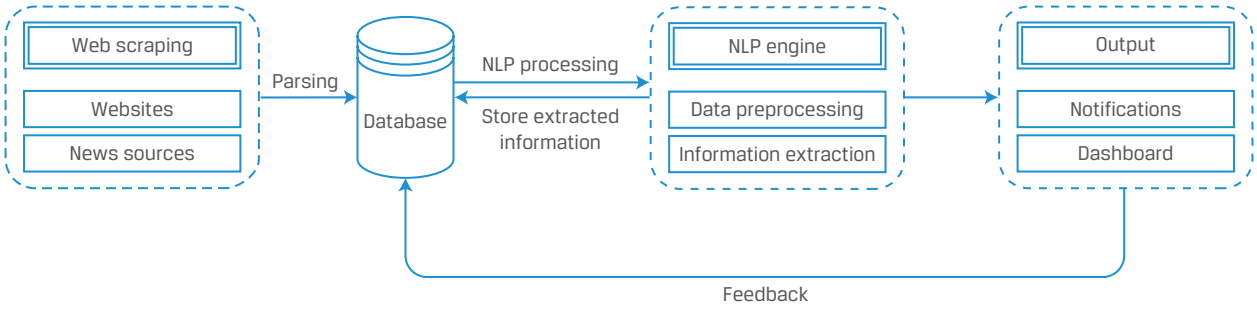


Exhibit 20. Sample Code: Identifying Entities within Documents

Import packages and the model

```
import spacy
import en_core_web_trf
```

```
spacy_model_name = 'en_core_web_trf'
if not spacy.util.is_package(spacy_model_name):
    spacy.cli.download(spacy_model_name)
nlp = spacy.load(spacy_model_name)
```

Identify entities in text

```
text = "Elon Musk pulls out of $44bn deal to buy Twitter."
doc = nlp(text)
```

```
for entity in doc.ents:
    print(entity.text, entity.label_)
```

```
Elon Musk PERSON
$44bn MONEY
Twitter ORG
```

the company "State Street." These limitations encourage researchers to further fine-tune and improve their NER models. One approach is to leverage threshold-based matching techniques to achieve the desired accuracy. By increasing the threshold to accept matches using NLP-based similarity metrics, we can improve the accuracy of the models. However, this increased accuracy comes at a cost: We may have more false negatives because our model may be less forgiving of misspellings and short names. As a result, researchers should assess the trade-offs between false positives and false negatives in their applications.

Another approach to improve NER involves applying additional code to the results from the spaCy model. This allows more flexibility for researchers to customize their requirements and insights for identifying entities. This approach may be helpful to deal with the difficulties in identifying such companies as AT&T, where its ticker (T) and acronyms are quite common in normal language usage.

Overcoming Business Challenges

On top of these technical challenges, various business issues slow or hinder the adoption of NLP across asset managers. In our view, the biggest hurdle is data—data quality and accessibility. Until recently, data had not been viewed as an asset, and as a result, the quality of data is varied and in many cases, unknown. We see many companies creating data organizations to tackle these issues, with the emergence of the chief data officer as an important executive and driver of business prioritization.

With the appropriate focus and resourcing, firms can leverage data as an asset, feeding data into dashboards

and models that ultimately help with faster and better decision making. This is particularly true for NLP pipelines and tools because the existing policies around these data are likely nonexistent, which presents an opportunity for organizations to design governance processes from scratch. In addition, firms will need to be nimble in their governance and policies because of the nascent and evolving regulatory frameworks; regulators are struggling to keep up with the breadth and complexity of data and the uses of data across commercial platforms, but we expect improved guidance in the future.

Data accessibility is also a big hurdle: Data scientists face challenges in finding and ingesting the required data for analysis and modeling. With regard to external data, asset managers can choose to create their own datasets or partner with vendors. Our view is that the most innovative and impactful data sources will be harder to obtain, and as a result, asset managers will need to develop their own pipelines to ingest these data. This may involve web scraping and capturing various forms of communication (html, different text file formats, audio, and video). Fortunately, many open-source tools exist to handle these situations.

As datasets become widely used and commoditized, vendor solutions are likely feasible and cost effective. For example, about 10 years ago, asset owners creating NLP signals from corporate filings, such as 10Ks and 10Qs, and earnings call transcripts developed pipelines to scrape, clean, and ingest the documents directly from the SEC's EDGAR website. Today, various vendors offer data solutions that significantly simplify the pipelines to leverage corporate filings. Another example is ESG data: Currently, there is little consistency in ESG metrics, standards, reporting formats, and disclosure frequencies. This means that asset managers are developing bespoke NLP pipelines to process the various documents and metrics on company websites, disclosures, and presentations. Over time, requirements will evolve because we expect this space to consolidate considerably over the next few years and vendors will emerge to develop solutions around ESG data collection, analysis, and reporting.

Transforming Culture

An organization's culture may be another important hurdle in its transformation. Organizations aspire to learn, innovate, and adapt, but cultural norms and human behavior can sometimes impede progress. We see this repeatedly across the industry.

Successful portfolio managers at all organizations typically have successful track records and significant assets under management from clients who have bought into their capabilities. Although these portfolio managers strive to outperform and deliver for their clients, one clear tendency is for teams to continue doing what has been successful. Investment teams operate on clearly articulated investment philosophies that have been honed over many years. They apply these philosophies using disciplined investment processes to uncover insights and opportunities. These structures have brought the team historical success, so changing or overriding these processes will be extremely difficult. This is the challenge that new data sources and more advanced techniques, such as NLP, face—overcoming the entrenched mindset that may come with historical success. This issue is compounded because clients expect investment teams to continue with processes that have worked and brought them success. However, with the financial markets, competitors, data sources, and investment tools changing rapidly, successful teams will need to evolve to survive.

To help organizations in their transformation, data scientists should leverage techniques to augment rather than replace existing processes, which is best achieved through pilot projects focused on real problems. Projects should be driven by investment controversies, sales opportunities, or operational gaps. By partnering with the decision makers, data scientists can achieve shared outcomes that are valued by all parties, thus ensuring adoption and success. NLP projects focused on addressing inefficiencies, such as gathering data, updating tables, and systematically making data more accessible, can deliver value and encourage adoption. Our use cases discussed in the earlier sections are examples where we have seen success because they address existing gaps that can be addressed through data science techniques.

Once investors, sales teams, and operations functions understand the value and capabilities of data science, they will be more open to further exploration. For example, once the investment teams have access to the corporate filings on a common platform (such as a dashboard), it is natural to extend the capabilities to synthesize the documents. Sentiment analysis and topic extraction are examples of common extensions once the data are in place. Indeed, to ensure adoption and success, asset managers should engage with end users to implement these ideas. We find that summarization, topic modeling, and question answering projects have the highest probability of success because they address common inefficiencies or manual processes.

Asset managers may see opportunities to leverage NLP to achieve efficiencies in operational processes. Indeed, as we discussed in our use cases, our techniques can be applied across various parts of the organization. While this may be low-hanging fruit, cultural norms and mindsets may make it difficult for innovative ideas to be adopted quickly. Employees may be concerned about their roles and livelihood when new technology is introduced: What is their place once the new techniques are incorporated? What incentive do they have to adopt the new technologies and potentially

eliminate the need for their roles? As a result, asset managers need to assess cultural barriers when trading off the operational savings versus the likelihood of success.

Developing Clear Success Metrics

Companies of all stripes are intrigued by the potential of new data, new techniques, and new technologies in their organizations. However, it is important to define clear success metrics on these new initiatives. For organizations starting their exploration, modest goals are likely appropriate. For example, getting one investment team to adopt and thus champion the successes may be a realistic and achievable goal. For more advanced organizations, developing common infrastructure and tools while nurturing talent may be the appropriate goals. In all cases, however, we suggest organizations take long-term views on their projects and articulate short-term milestones that create awareness, maximize adoption, and develop competencies.

Attracting and Developing Talent

There will be a talent gap. Existing employees may not possess the required data science or NLP skills, so companies will need to hire external talent. Creating a "Center of Excellence" that aims to raise the capabilities across the organization is one way to bridge the talent gap. While exact structures may differ from firm to firm, we believe it is important to have a dedicated centralized team that keeps abreast of industry developments, creates new capabilities, shares best practices, builds common infrastructure, and develops talent. Achieving these outcomes will be maximized with a centralized team with the mandate to raise the competencies of the entire organization.

This central hub acts as a critical link among all the teams in the organization, ensuring that the latest insights and capabilities are shared broadly. With this core team in place, organizations now have different approaches to build out their data science capabilities. On the one hand, augmenting the centralized team with embedded data scientists within business functions ensures that there is domain expertise and the business units are accountable for the success of the teams. On the other hand, consolidating all the data scientists in one centralized team ensures there are efficiencies and few overlaps on projects, but the researchers may be more removed from the end users. Ultimately, the optimal structure will depend on the organization's existing norms, culture, and goals, but having a Center of Excellence is essential to long-term success.

The Road Ahead

NLP has tremendous potential in asset management. Our use cases highlight existing areas where NLP is already having an impact. We expect growing adoption across more functions in asset management organizations as three trends take hold: (1) availability of and improvements in open-source models, training data, tools, and vendor options, (2) fine-tuned models with proprietary insights, and (3) development of non-English capabilities.

Improved Tools

Even though there were over 70,000 models and counting as of late 2022 on Hugging Face (a popular NLP platform), we expect to see continuous innovation in training and fine-tuning techniques. Models will be trained for specific tasks, thus improving their performance across various activities. For example, customized BERT models may be developed for different fixed-income sectors, such as collateralized loan obligations, mortgages, private credit, and municipal bonds, among others. Semantic improvements will also be made, as machines are taught to further understand the nuances of language.

Asset managers are also exploring the use of audio and visual cues to augment the text-driven features we have discussed. By combining the tonal variations of spoken text, the facial and bodily expressions of the speakers, and the actual text used, researchers hope to have a more comprehensive understanding of the intended communication. Indeed, in our day-to-day interactions, we incorporate all these elements (tone, delivery, body language, and content) to discern the message and its nuances. Interestingly, techniques such as the ones discussed in this chapter can be extended to capture audio and visual features. For example, different aspects of the audio message can be embedded using BERT-type models, and these embeddings can be compared and manipulated to tease out changes in emotions and reactions.

We have seen an explosion in the breadth of vendor offerings to solve specific NLP tasks in investment, compliance, and operational activities. Specifically, document processing such as that in our use cases on theme searches and question answering is becoming more mainstream. Vendors have developed customized solutions for specific applications across the industry (ESG documents, regulatory filings, shareholder documents, etc.). While customized solutions are the natural starting points, we expect more consolidation to occur and more generalized and robust solutions in the future. For example, QA techniques will be more powerful and able to handle diverse types of questions across different documents. To draw an analogy from another domain, the Google search engine has evolved over time and is now able to handle specific queries rather than simply return generalized results. Searching for the "score from yesterday's Yankees game" returns the actual score rather than links to the team website or a list of baseball scores. We expect these capabilities to materialize for the asset management industry, giving researchers the ability to answer specific questions and synthesize ideas quickly.

Proprietary Models

We have seen various instances of asset owners injecting their own insights into language models to improve the base models. There are two common approaches—proprietary dictionaries and proprietary phrase annotations. As discussed earlier, various public dictionaries exist to score sentiment on sentences and documents. Having said that, investment teams can create custom dictionaries to score documents based on their own preferences, thereby tailoring the NLP algorithms for their specific use cases. This customization enables asset managers to synthesize documents using their own views and insights and build trust in the underlying algorithms.

Beyond customized dictionaries, investment teams can also label or annotate words, phrases, and sentences to improve language models, such as BERT. Our own research suggests language models can be more accurate on sentiment classification and other NLP tasks with feedback from human annotators. The following two sentences are from a telecom company that FinBERT scores as positive.

Sentence 1: "It does seem as though your competitors are healthfully ramping up their efforts to try and improve service quality."

Sentence 2: "Churn in handsets remains below 0.9%, but it sure looks like it was up a lot year over year."

From a sentiment perspective, we view these sentences as negative when the broader context is considered. In the first sentence, a company's competitors performing better is typically negative for the company, but FinBERT focuses on the "ramping up" and "improve service quality" to determine its positive sentiment.

As for the second sentence, higher "churn" for a telecom company should be viewed as a negative even though such models as FinBERT normally view "up a lot year over year" as positive. These examples illustrate the need for language models to be further trained on specific domains for them to be effective.

By manually labeling these sentences and using them to fine-tune FinBERT, we found marked improvement in the model's performance. We expect more firms to use manual annotations to improve the base models and to engender more trust in the models.

While we largely used the investment arena for our discussion on proprietary dictionaries and phrases, the same concepts can be leveraged in other parts of the organization. Sales teams can customize NLP tools for their client outreach using the same methods to optimize client interactions and sales effectiveness.

Language Expansion

Finally, we expect to see substantial progress in the development and advancement of non-English language models. While many of the NLP developments started with English documents, there has been both academic and practitioner interest in other languages. Specifically, as firms look for an edge in NLP, one fruitful path may be using documents in local languages rather than using the translated English versions. Local documents may reveal tone and semantic references that are discernible only in the original language.

Conclusion

As NLP tools mature in the asset management industry, organizations will be able to apply these tools to a wide range of problems. We are already seeing early successes in investments, distribution, and operations, and we expect this momentum to continue. As decision makers become more comfortable with the new tools and models, adoption will accelerate and efficiency gains will accrue quickly. Over time, with appropriate engagement and training by humans, the NLP tools will become more transparent and effective and users will readily incorporate these tools into their arsenal.

Ultimately, these NLP tools improve our decision making, and that alone should ensure their adoption and success.

References

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. "Language Models Are Few-Shot Learners." Cornell University, arXiv:2005.14165 (22 July).

Cohen, Lauren, Christopher J. Malloy, and Quoc Nguyen. 2019. "Lazy Prices." Academic Research Colloquium for Financial Planning and Related Disciplines (7 March).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." Cornell University, arXiv:1810.04805 (11 October).

Grootendorst, M. 2022. "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure." Cornell University, arXiv:2203.05794 (11 March).

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9 (8): 1735–80.

Huang, Allen, Hui Wang, and Yi Yang. Forthcoming. "FinBERT: A Large Language Model for Extracting Information from Financial Text." *Contemporary Accounting Research*.

Hutto, Clayton J., and Eric Gilbert. 2014. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media* 8 (1): 216–25. https://ojs.aaai.org/index.php/ICWSM/article/view/14550/14399.

Joshi, Mandar, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. "SpanBERT: Improving Pre-Training by Representing and Predicting Spans." *Transactions of the Association for Computational Linguistics* 8: 64–77.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." Cornell University, arXiv:1907.11692 (26 July).

Loughran, Tim, and Bill McDonald. 2011. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *Journal of Finance* 66 (1): 35–65.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." Cornell University, arXiv:1301.3781 (7 September).

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. "GloVe: Global Vectors for Word Representation." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (October): 1532–43.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." Cornell University, arXiv:1910.10683 (28 July).

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. "SQuAD: 100,000+ Questions for Machine Comprehension of Text." Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (November): 2383–92.

Turing, Alan Mathison. 1950. "Computing Machinery and Intelligence." *Mind* LIX (236): 433–60.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 5998–6008.