HANDBOOK OF ARTIFICIAL INTELLIGENCE AND BIG DATA APPLICATIONS IN INVESTMENTS

II. NATURAL LANGUAGE UNDERSTANDING, PROCESSING, AND GENERATION: INVESTMENT APPLICATIONS

This book can be found at cfainstitute.org/ai-and-big-data

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

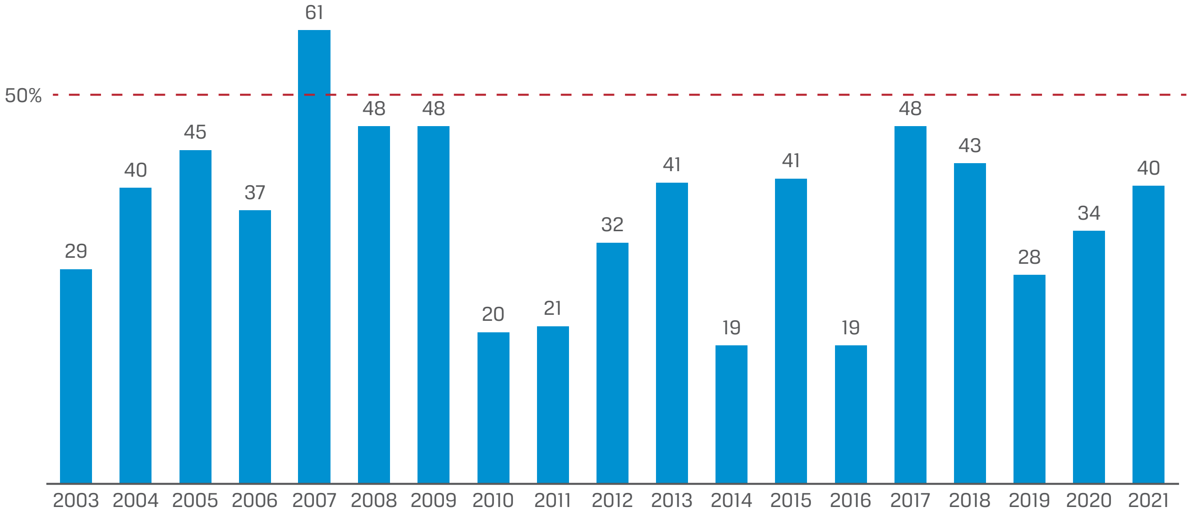
While investors are searching for higher returns, active managers, overall, have not delivered on their promises. Over the past decade, active managers have underperformed their respective benchmarks. **Exhibit 1** shows the percentage of US large-cap equity managers outperforming the broad US market 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

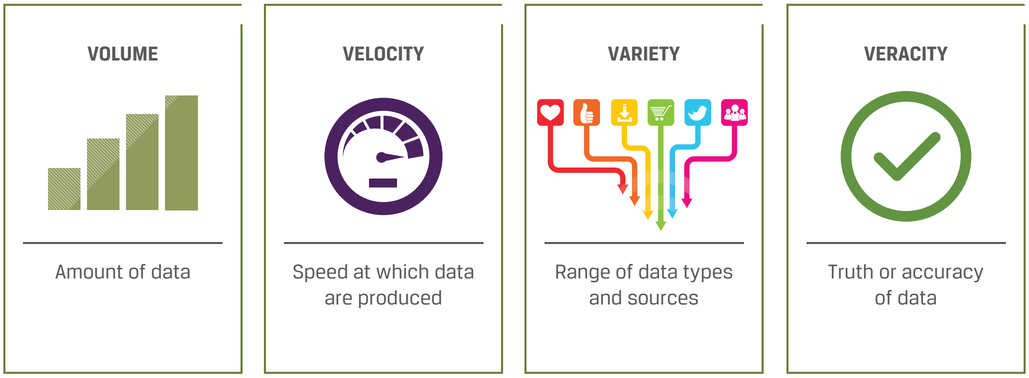
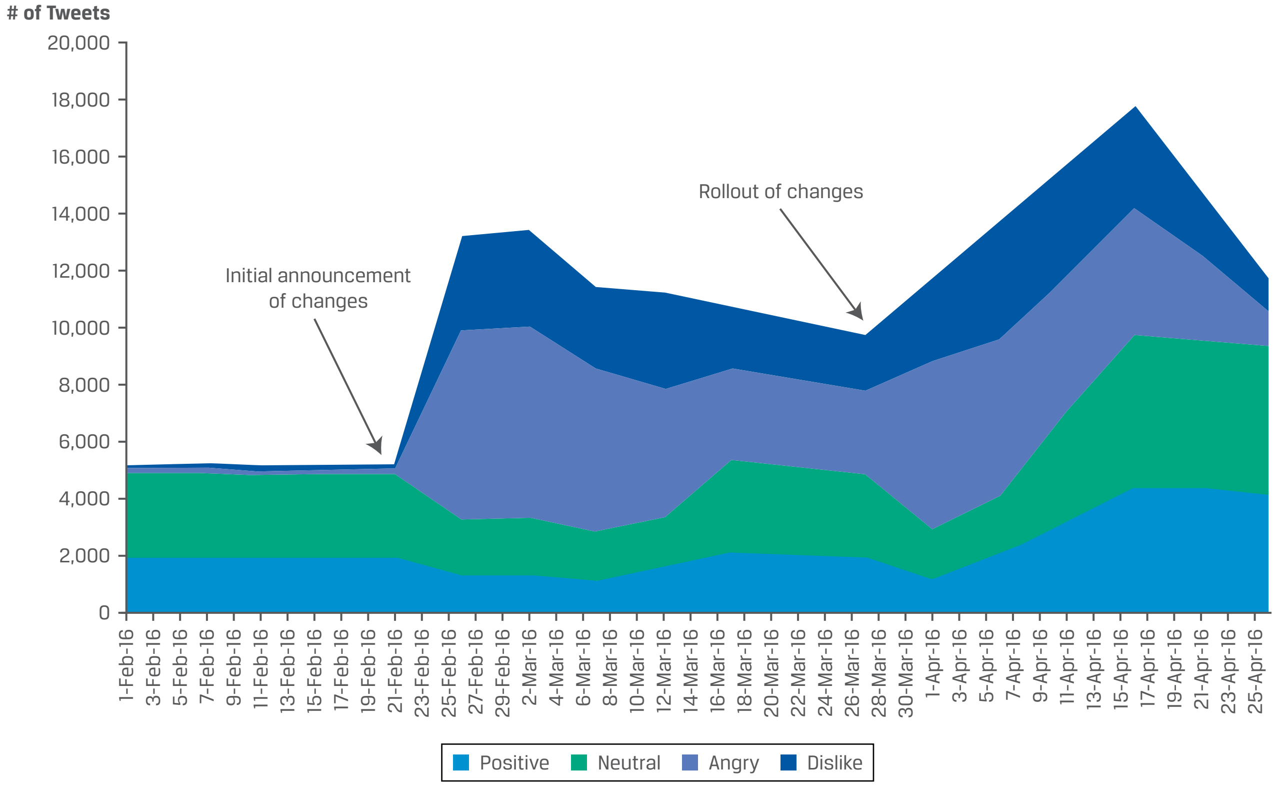

VOLUME
Amount of data
VELOCITY
Speed at which data are produced
VARIETY
Range of data types and sources
VERACITY
Truth or accuracy of data

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 rollouts 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

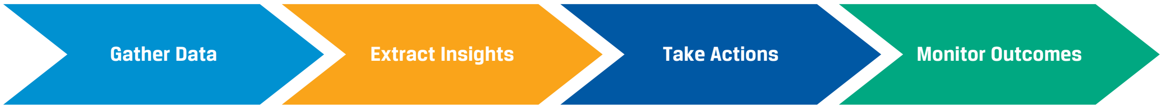
In this section, we discuss how NLP is used in finance and asset management.

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.

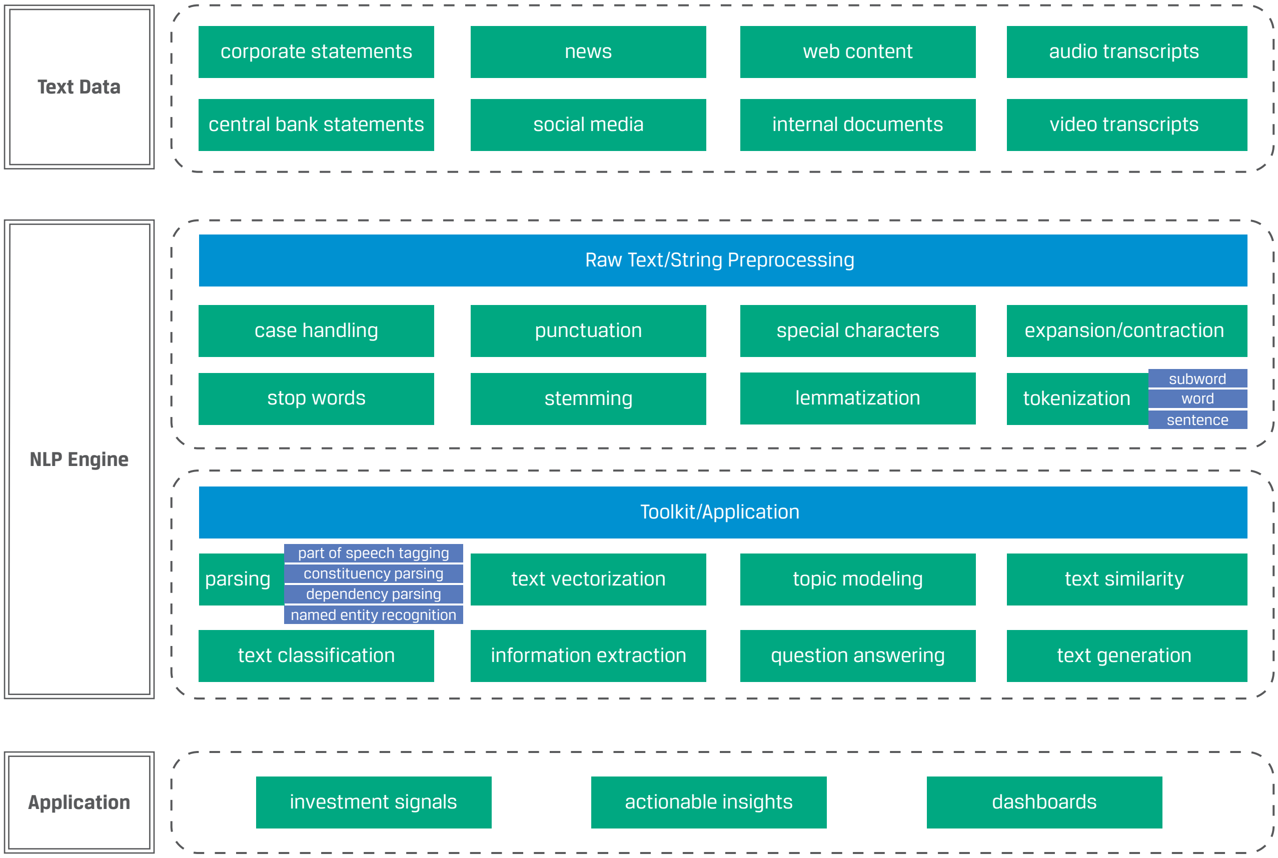
Typical NLP Pipeline

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.

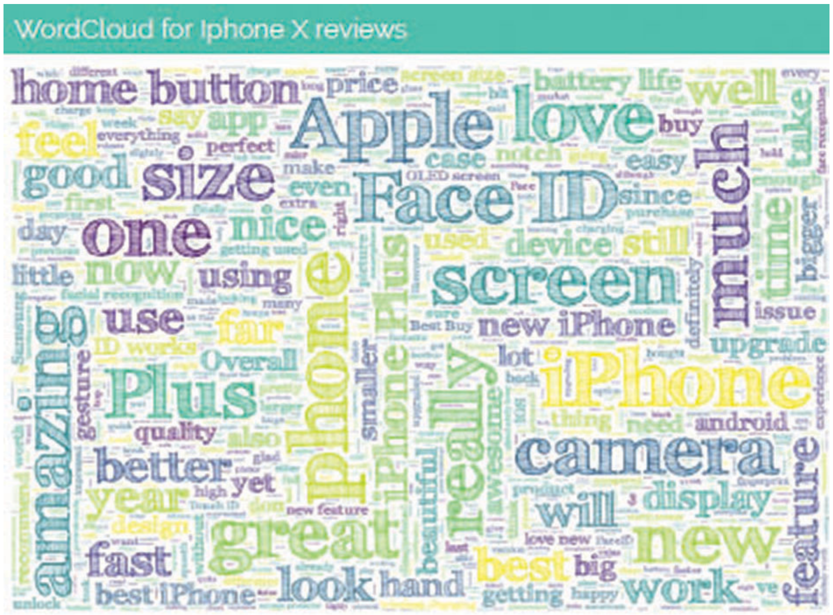
The larger words represent the most frequently used words in the reviews. While some words are meaningless ("much" and "use") or expected ("Apple"), it may have been reassuring to see that "Face ID" was being highlighted and such words as "amazing" and "love" were being used to describe the new phone.

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.

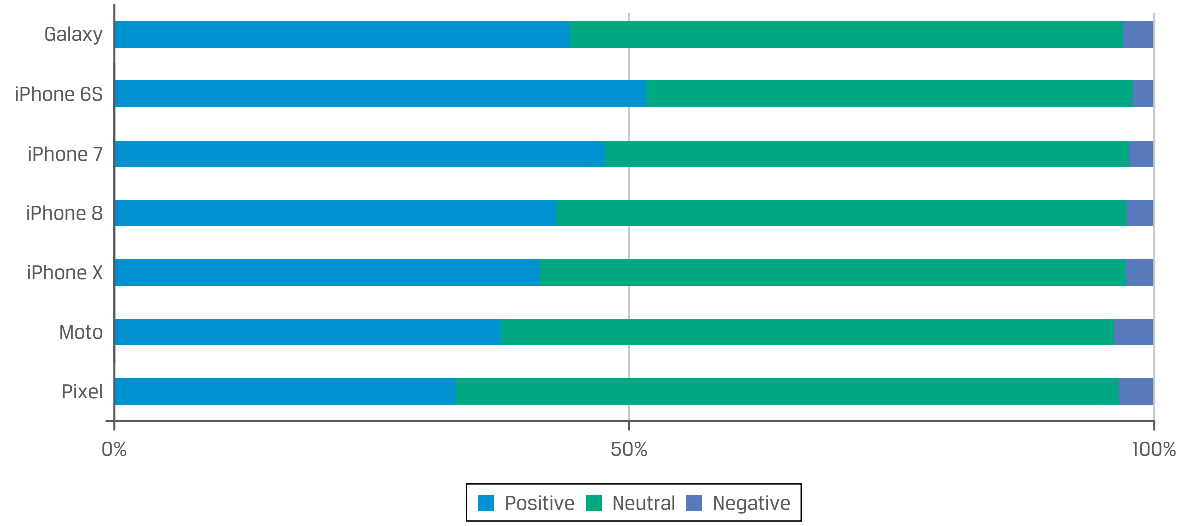
This simple example illustrates the basics of NLP analysis. Once the raw data from social media posts and customer reviews are ingested, simple tools can be used to analyze the data to provide insights into investment controversies. In late 2017, investors were assessing the launch of the iPhone X, and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

Exhibit 6. Word Cloud for iPhone X Reviews



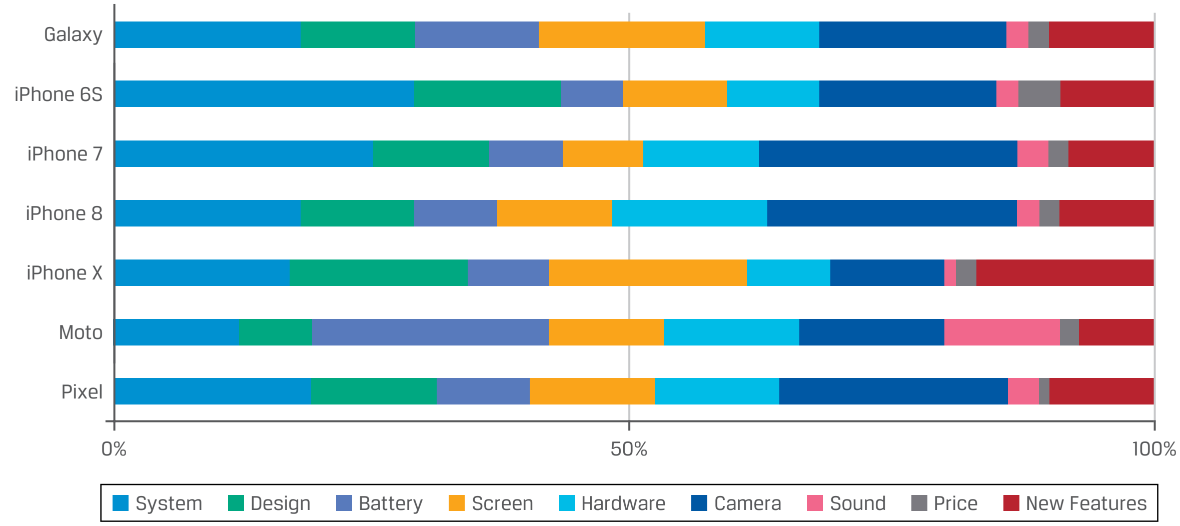
Source: AllianceBernstein.

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.

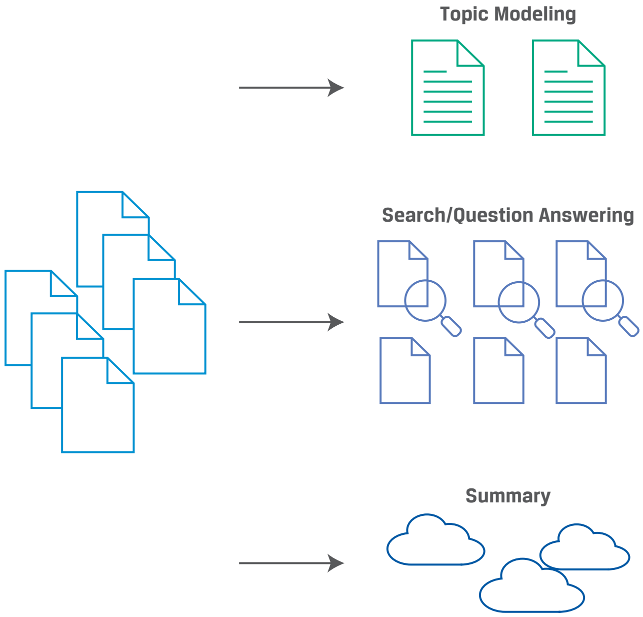
Extracting Themes to Highlight Important Topics and Trends

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

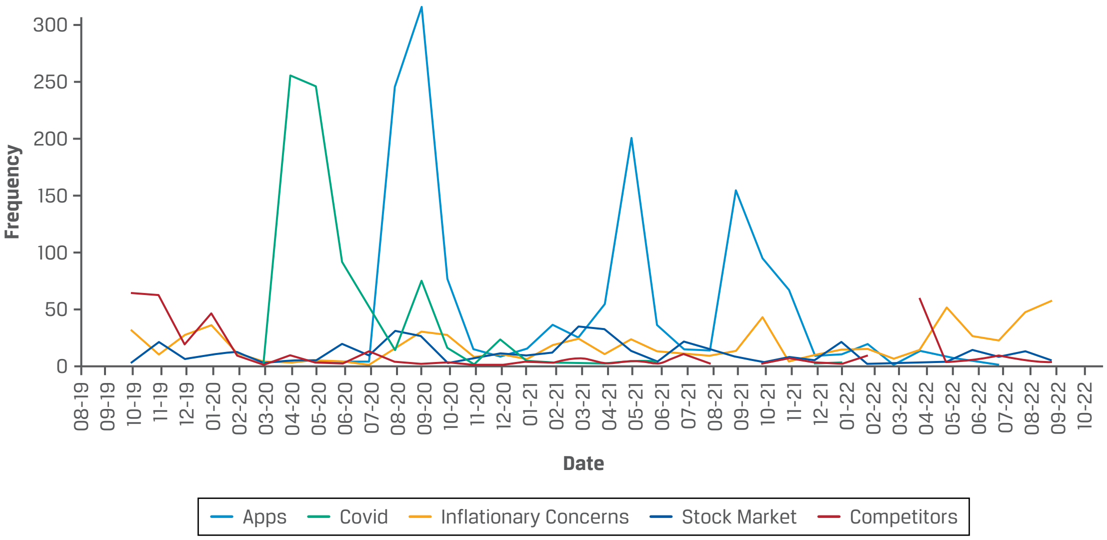


representation, clustering the documents, and then extracting keywords to represent the derived themes.

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

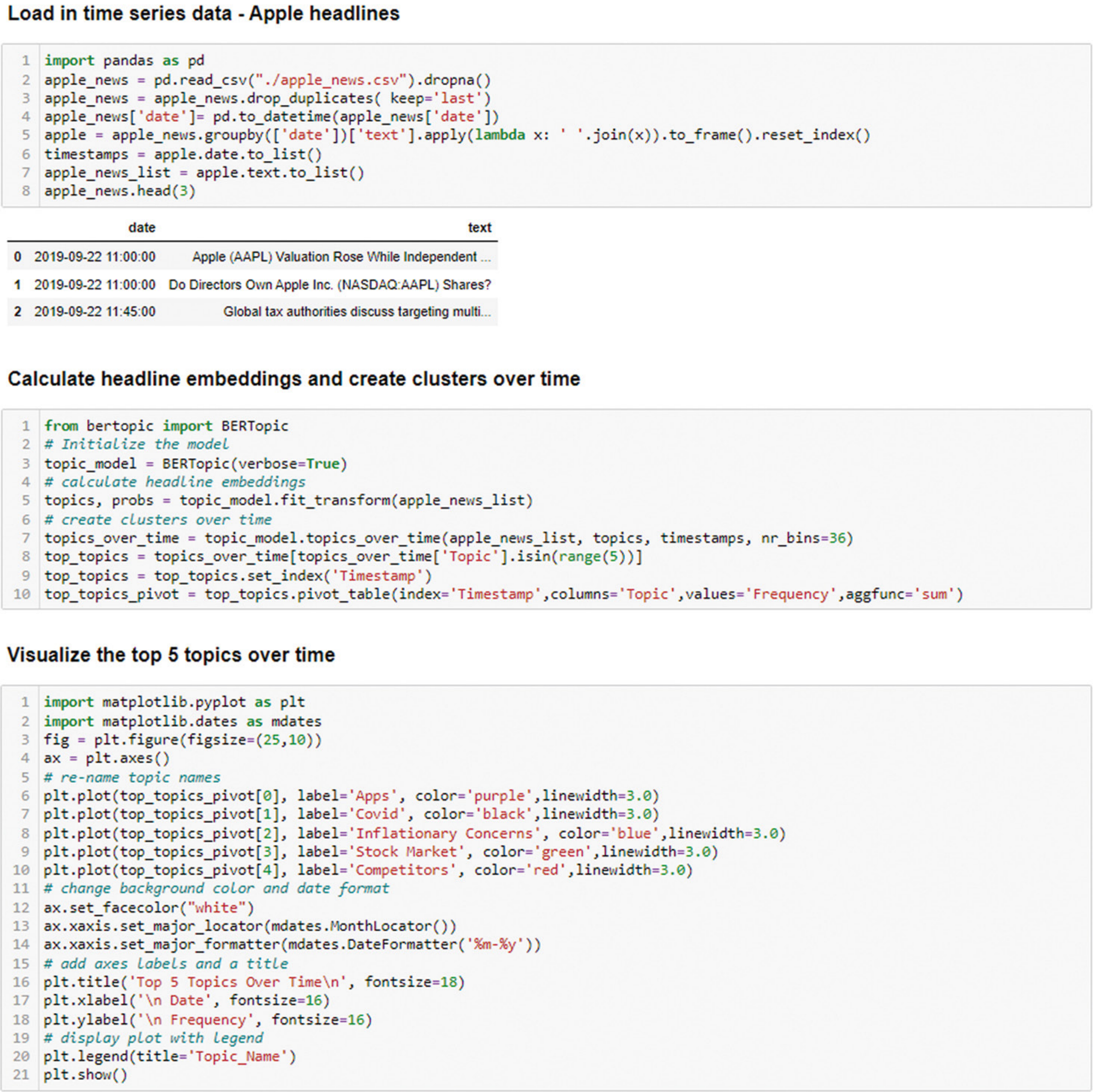


Source: AllianceBernstein.

"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

```python
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

```python
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

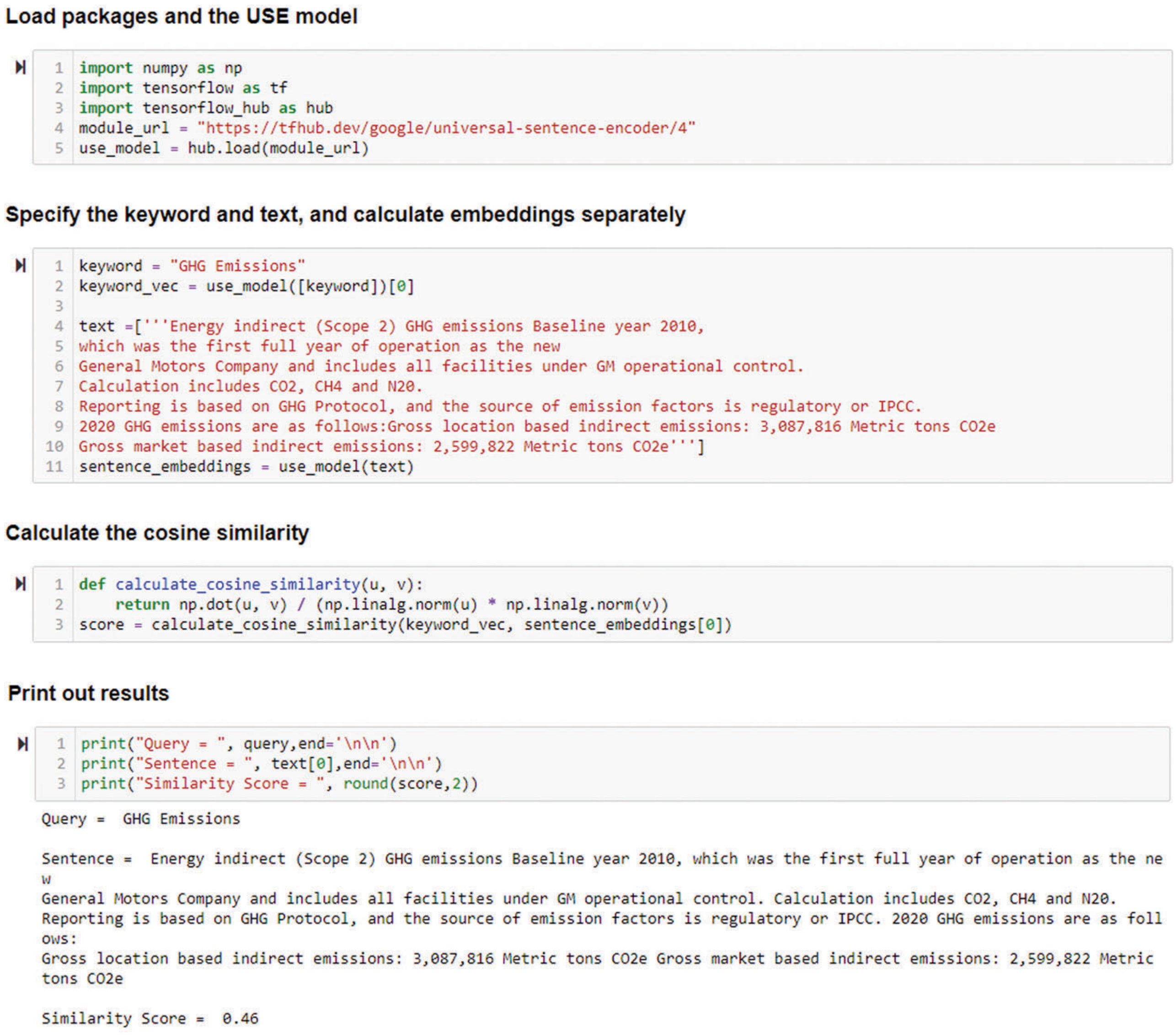
```python
import matplotlib.pyplot as plt
import matplotlib.dates as mdates
fig = plt.figure(figsize=(25,10))
ax = plt.axes()
# re-name topic names
plt.plot(top_topics_pivot[0], label='Apps', color='purple',linewidth=3.0)
plt.plot(top_topics_pivot[1], label='Covid', color='black',linewidth=3.0)
plt.plot(top_topics_pivot[2], 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))
```

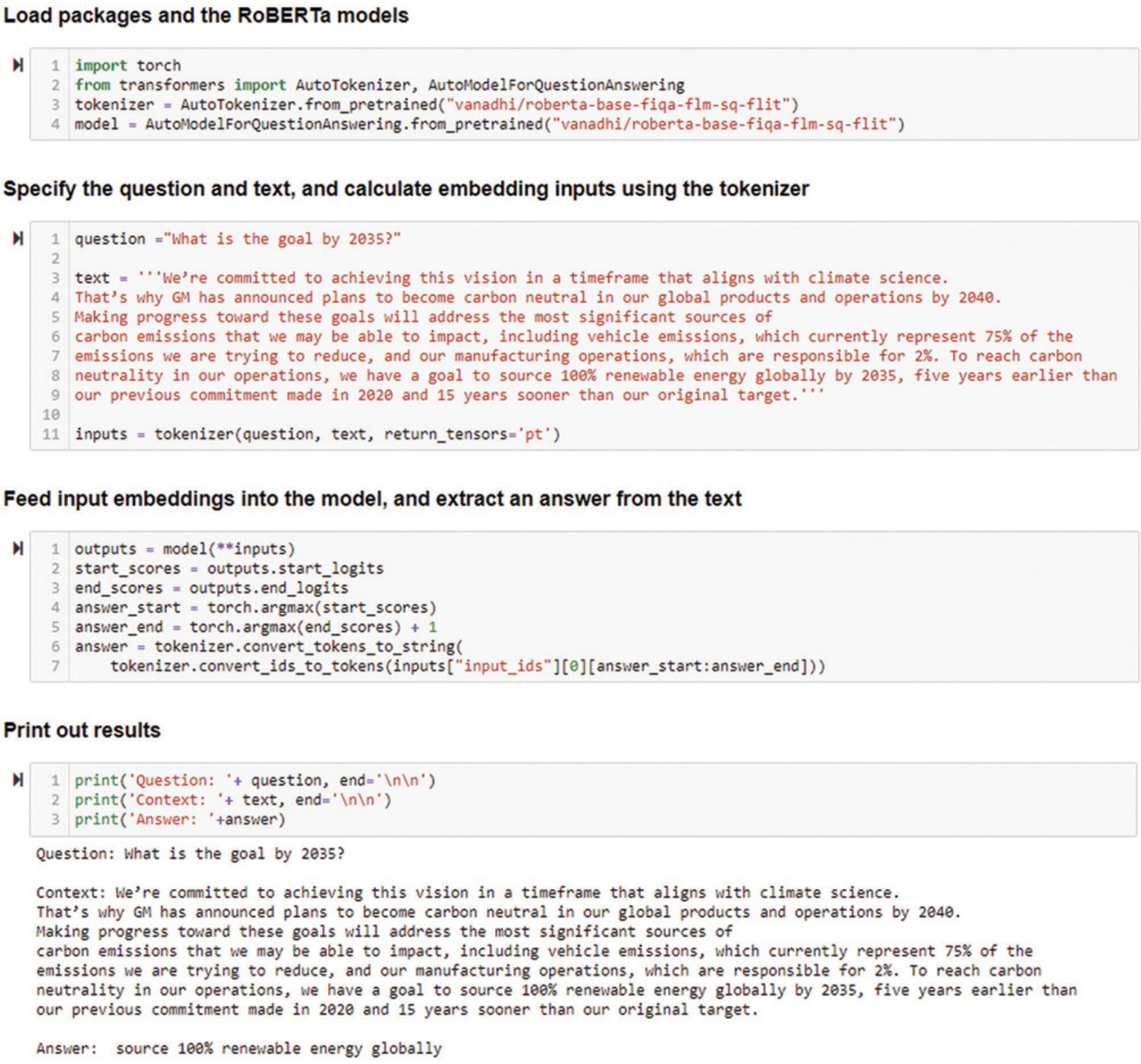
```
Query =  GHG Emissions

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

```python
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

```python
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

```python
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

```python
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.

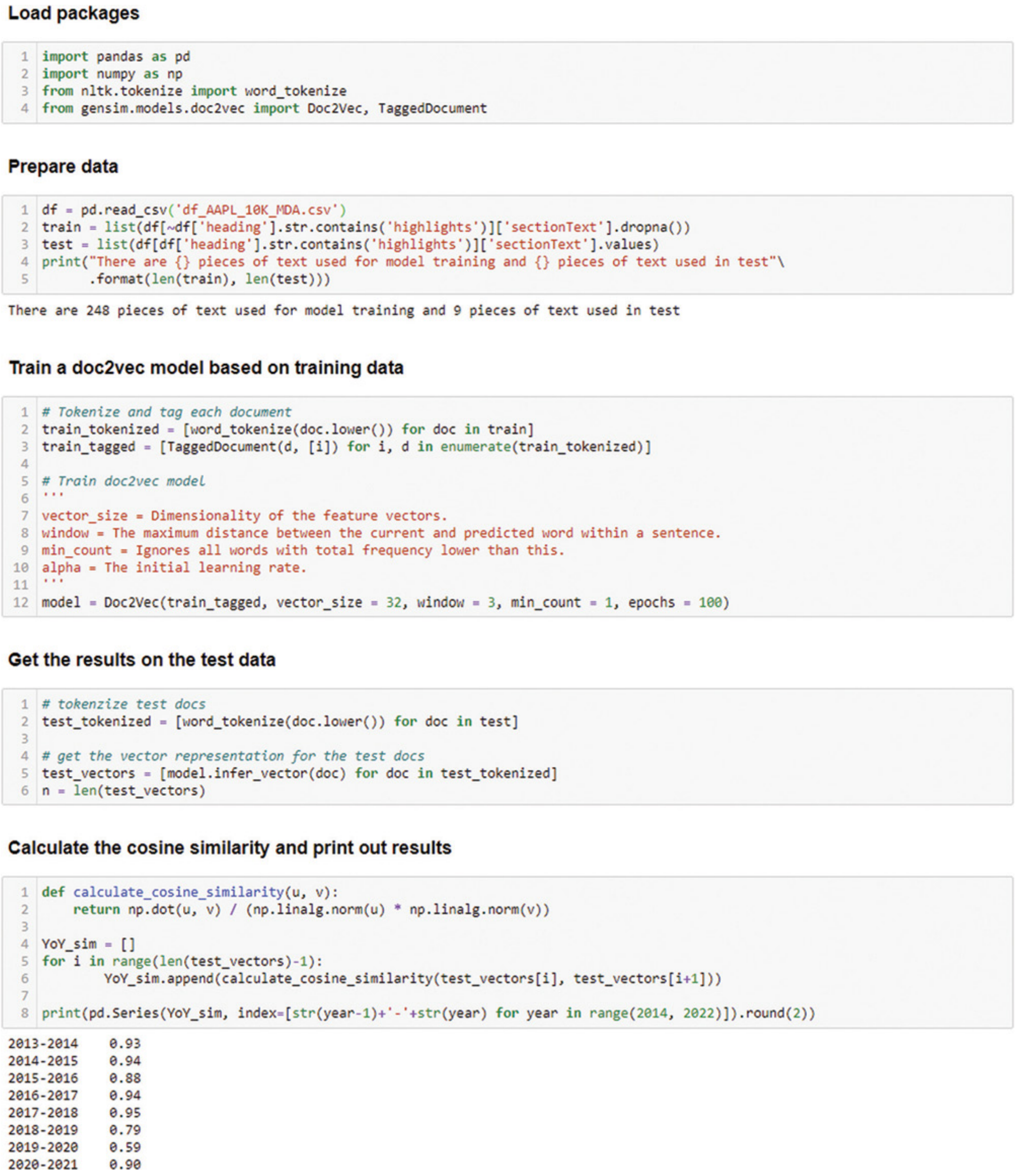
A variety of NLP approaches can be used to generate investment signals from earnings call transcripts. Bag-of-words approaches using predefined dictionaries and context-driven language models are common techniques. We describe three categories of features in our analysis of the transcripts—document attributes,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

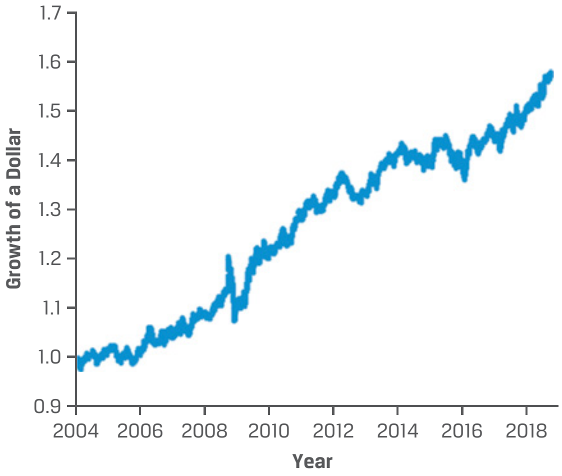
Calculate the cosine similarity and print out results

```
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.

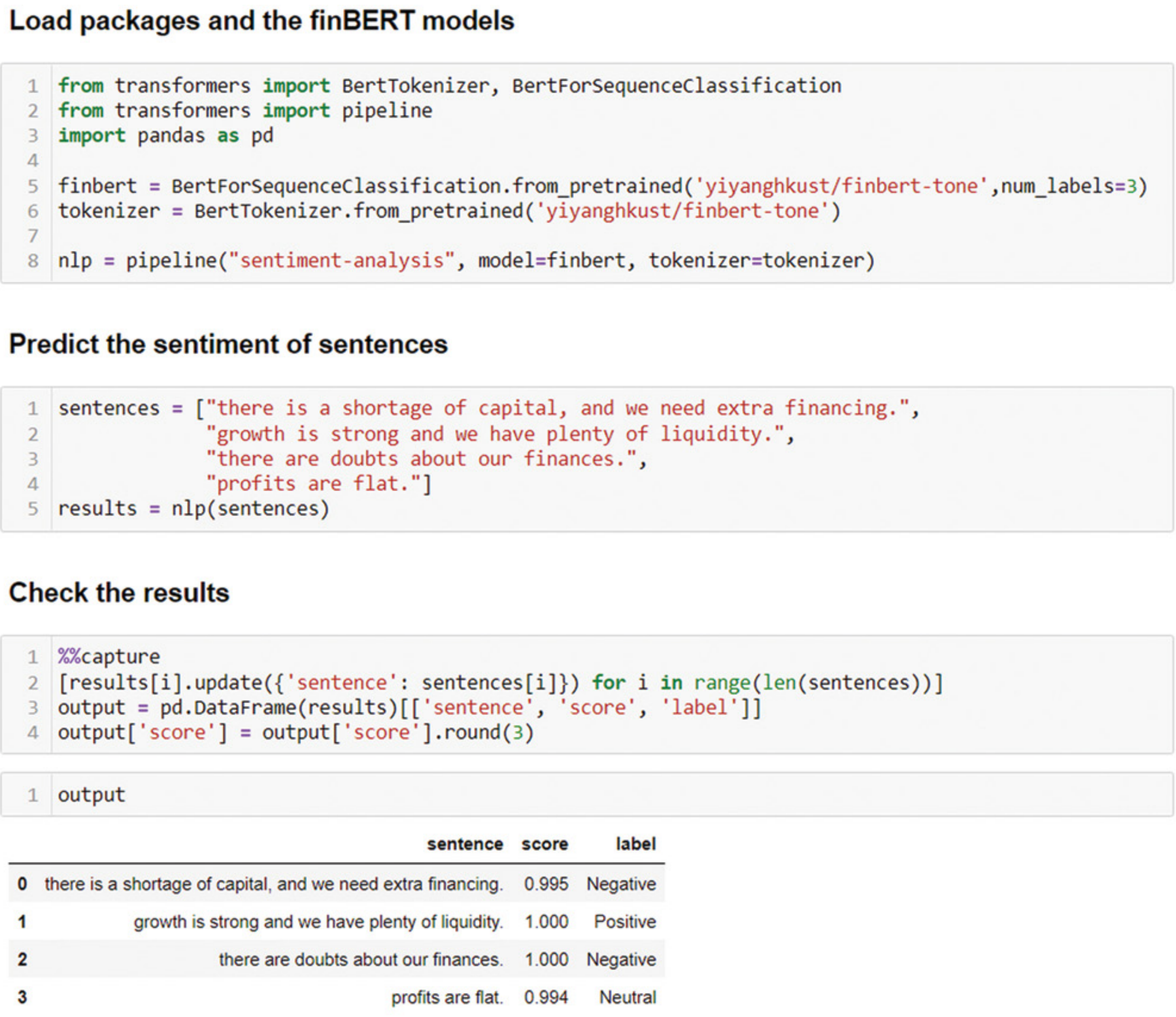
Exhibit 19 illustrates this process. First, public data can be obtained through web scraping. For example, data scraping pipelines can be built to detect document changes on various websites. We can also monitor major news outlets to track mentions of specific keywords, such as names of organizations, 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.

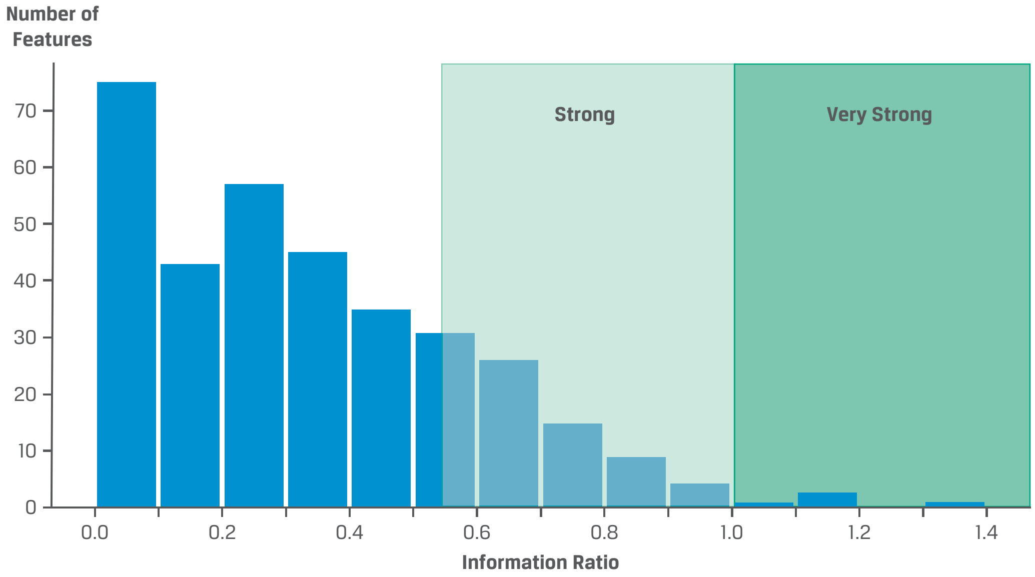
Overcoming Technical Challenges

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.

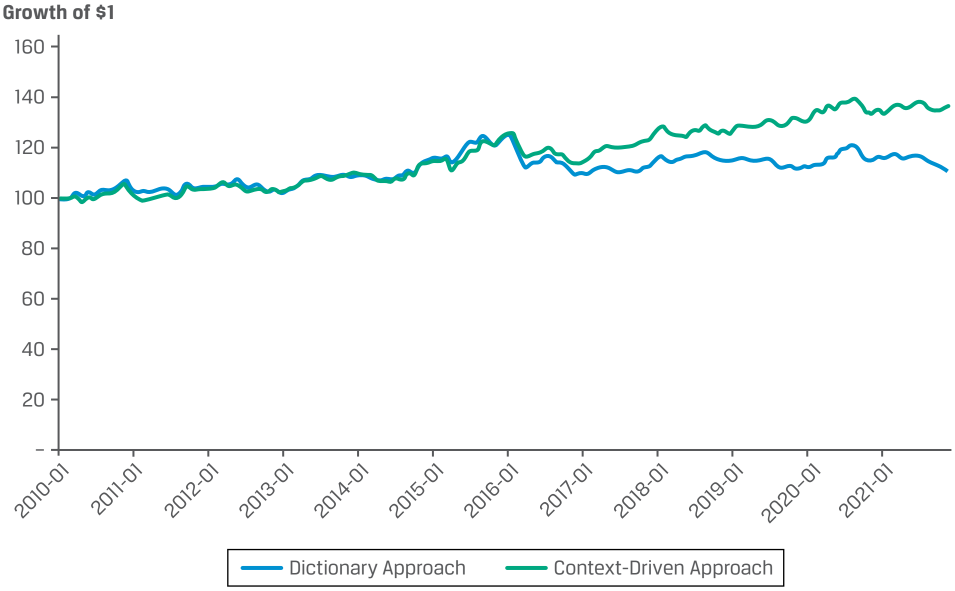
Even context-specific language models, such as BERT, will need to learn from new financial domains. While FinBERT-type models have been trained on financial text, further training and fine-tuning is likely required to improve

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

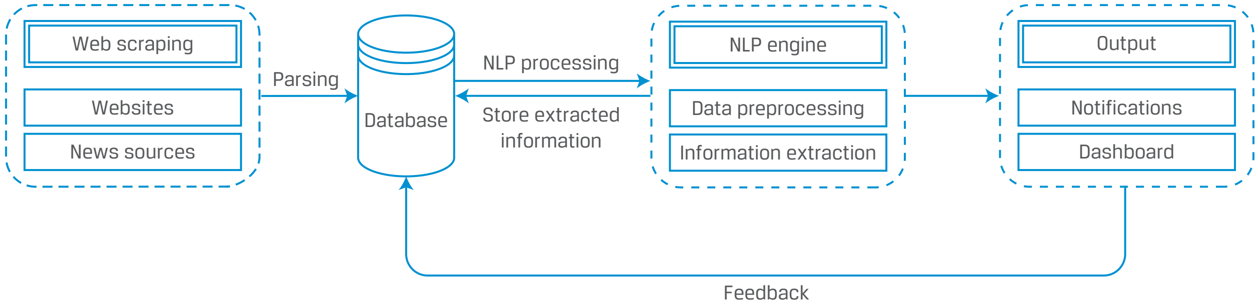
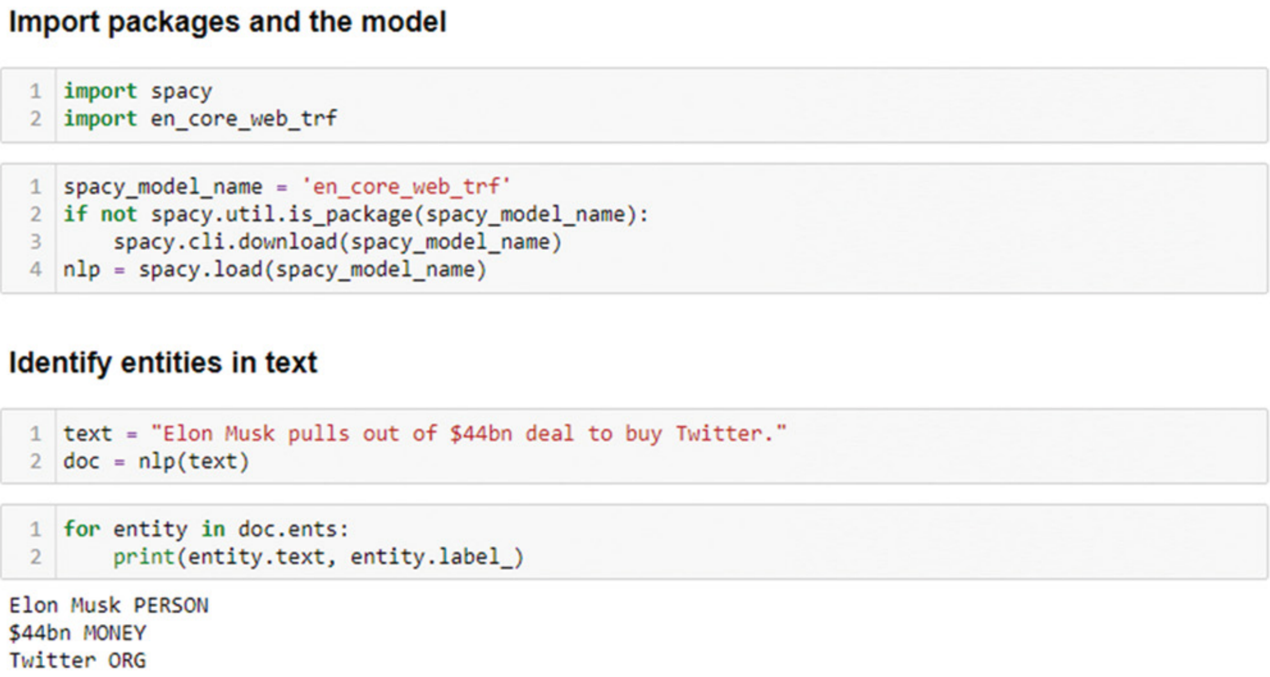


Exhibit 20. Sample Code: Identifying Entities within Documents



Import packages and the model

```python
import spacy
import en_core_web_trf
```

```python
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

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

```python
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.

5. ADVANCES IN NATURAL LANGUAGE UNDERSTANDING FOR INVESTMENT MANAGEMENT

Stefan Jansen, CFA
Founder and Lead Data Scientist, Applied AI

Introduction

Investors have long sifted through documents searching for facts that corroborate fundamental analysis, support forecasts, or provide insight into market dynamics. Given the importance of language for communication, it is not surprising that 80%–90% of information is available only in unstructured forms, such as text or audio, making it much harder to work with than tabular data that conform to database and excel formats (Harbert 2021). However, such data can enable more valuable, differentiated insights than numerical data in widespread use alone.

Text data sources relevant for investors depend on the asset class and investment horizon. They can be very diverse, ranging from regulatory filings and research reports to financial news and social media content. In addition, legal contracts may play an essential role for, for example, debt or real estate investments. Substantial improvements in speech recognition enable real-time processing of audio data, such as earnings calls, adding speaker tone, and other verbal nuances, which may uncover hidden insights. Similar advances in machine translation expand this universe to any common language. Over the last two decades, *electronic access to text data* with information affecting investment decisions in real time has increased by orders of magnitude. Automated downloads of corporate 10-K and 10-Q SEC filings, for instance, multiplied almost 500 times, from 360,861 in 2003 to around 165 million in 2016 (Cao, Jiang, Yang, and Zhang 2020). Machines have been driving part of this data explosion: In 2015, the Associated Press (AP) was already generating over 3,000 stories automatically on US corporate earnings per quarter, a tenfold increase over previous human production. The output had grown by more than 50% by 2017, while other outlets followed the AP's lead.

Unsurprisingly, the resulting data deluge has accelerated research into *natural language processing* (NLP), the AI discipline that cuts across linguistics and computer science and focuses on the ability of machines to make sense of text data. Practical goals are to automate and accelerate the analysis of text data or augment the capabilities of humans to do so. In 2016, J.P. Morgan deployed technology that could read 12,000 documents to extract 150 attributes relevant for human review in seconds, compared to 360,000 person-hours per year needed previously, all while being more accurate. NLP *use cases* vary from matching patterns and summarizing documents to making actionable predictions. They include standardizing and tagging investment targets across a large body of documents, scoring news sentiment or environmental, social, and governance (ESG) disclosures, flagging compliance issues, and forecasting market prices. *Sustained R&D* has improved applications that help manage and exploit the explosion of unstructured language data in versatile ways to create value along the investment process. These applications can boost analyst productivity by imposing structure on documents and their content, focus the attention of portfolio managers by prioritizing news flow, and facilitate risk management by detecting material events in real time. They can also support investment decisions by generating alpha signals.

The most impressive improvements build on *breakthroughs in machine learning (ML)*, particularly deep learning, where new language models have dramatically boosted performance on various tasks. However, organizations with a range of resources can leverage NLP applications because cloud-based application programming interfaces (APIs), high-quality open-source building blocks, and a growing vendor ecosystem have simplified the adoption or development of NLP solutions. At the other end of the spectrum, cutting-edge systems that generate proprietary alpha signals require an in-house team of specialists. This chapter describes NLP applications used by analysts and discretionary or systematic portfolio managers and characterizes emerging applications under development at leading quantitatively oriented funds. I begin by explaining key challenges facing NLP by machines and how recent advances have addressed them much better than earlier approaches. Then, I outline various categories of NLP applications and how they are used throughout the investment process. Finally, I illustrate three popular financial NLP use cases in more detail.

Challenges Faced by NLP

Making sense of language, never easy for machines, is even harder in finance.

Human language has many characteristics that make it *notoriously hard for machines* to process, interpret, and act on. It builds on complex grammatical rules, idiomatic usage, and contextual understanding—all of which humans take a significant amount of time to learn. Common challenges in interpreting language require decisions between multiple meanings for the same word, which may require context not provided in the document but instead assumed to be known to the reader. An early machine translation attempt illustrates how this process can go wrong: "out of sight, out of mind" translated into Russian and back to English produced "invisible, insane" (Pollack 1983).

A key challenge particular to machines is the requisite *conversion of text into numbers* suitable for digital computation. The output tends to simplify the intricate structure of language by omitting or weakening critical relationships among words, phrases, and sentences. As a result, algorithms have struggled for a long time to recover the meaning of statements and documents.

Popular applications such as sentiment analysis used relatively *simplistic, error-prone approaches*: They relied on the presence of specific words or language patterns defined by humans to evaluate statements or flag events, as we will see in more detail later. Such methods are transparent and intuitive but often have limited flexibility and sensitivity compared to language models with more sophisticated understanding.

Additional challenges emerge with *domain-specific uses of language*, as in finance. The need to account for the distinct semantics of technical terms and incorporate specialized vocabulary limits the transfer of applications that have proven helpful in other areas, such as online search, marketing, and customer service.

Sentiment analysis, for instance, one of the most popular financial NLP applications, is used across industries to evaluate customer feedback. However, as we will see in more detail, investor communication differs too much for results to carry over. These differences also limit the data that specialized applications can learn from. As a result, *text-based applications in finance are lagging* substantially behind developments in other domains.

Progress of NLP Applications in Investments

Notwithstanding these hurdles, ML applications for text data have advanced dramatically, driven by the familiar trifecta of more data, faster computation, and better algorithms. Paralleling critical developments in other AI disciplines over the last two to three decades, NLP has gradually abandoned techniques based on rules crafted by domain specialists, such as linguists, in favor of statistical models that *learn relevant patterns directly from text data*.

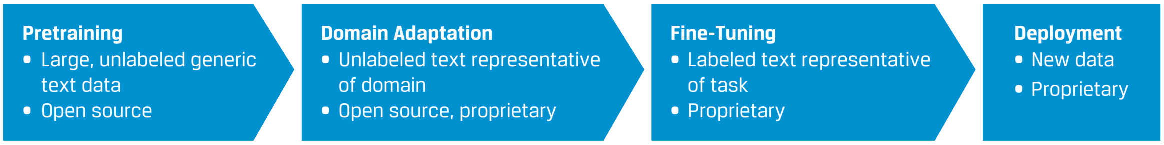
Over the last 5–10 years, language models based on novel deep neural network architectures have evolved to capture the nuances of language much more accurately than before. Recent breakthroughs involve using neural networks that learn how to *represent words as vectors* rather than individual integers (Mikolov, Sutskever, Chen, Corrado, and Dean 2013). These so-called embedding vectors contain hundreds of real numbers that reflect language usage in context because they are optimized to predict terms that are missing from a given piece of text. More specifically, the location of a vector captures much of a word's meaning in the sense of "know a word by the company it keeps" (Firth 1957). Not only are synonyms nearby, but even more complex relationships have vector equivalents: For example, the vector pointing from France to Paris would be parallel and of equal length to the one connecting Japan and Tokyo.

In another critical step forward, the *transformer architecture* abandoned recurrent neural networks, the hitherto dominant way to process text data due to its sequential nature, in favor of the attention mechanism (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin 2017). This innovation allows modeling dependencies between words regardless of distance and parallelizing the training process. The benefits are step changes in scale: Models can learn from much longer sequences, benefit from dramatically larger datasets, and rely on many more parameters.

State-of-the-art implementations now reach or exceed *human-level performance* on various benchmark tasks from sentiment analysis to paraphrasing or question answering (Wang, Pruksachatkun, Nangia, Singh, Michael, Hill, Levy, and Bowman 2019). Google introduced machine translation systems based on neural networks that reduced errors by 60% in 2016. Recent machine translation systems produce outputs that humans have difficulty distinguishing from those of a professional translator (Popel, Tomkova, Tomek, Kaiser, Uszkoreit, Bojar, and Žabokrtský 2020).

These large-scale models require a massive amount of unlabeled data and computation to extract their intricate understanding of language, which would make this new technology cost prohibitive for most applications. Fortunately, the performance breakthroughs reflect an emerging paradigm in developing AI systems toward a new class of *foundation models* (Bommasani, Hudson, Adeli, Altman, Arora, von Arx, Bernstein, Bohg, Bosselut, Brunskill, et al. 2021). This new model class learns universal capabilities from broad data (including but not limited to text data) that transfer to specific tasks very cost effectively. Instances of this new model class have begun to generate computer code and prove mathematical theorems (Polu and Sutskever 2020). Multimodal models that integrate, for example, text and audio data add intent parsing to speech recognition to capture nuances of language; under active research for voice assistants, applications aiming to read between the lines of earnings call transcripts are also emerging. Model performance appears to scale reliably with input data and model size, and the foreseeable

Exhibit 1. The Transfer Learning Workflow


Pretraining
• Large, unlabeled generic text data
• Open source
Domain Adaptation
• Unlabeled text representative of domain
• Open source, proprietary
Fine-Tuning
• Labeled text representative of task
• Proprietary
Deployment
• New data
• Proprietary

ability to boost both makes substantial further advances likely (Srivastava, Rastogi, Rao, Shoeb, Abid, Fisch, Brown, Santoro, Gupta, Garriga-Alonso, et al. 2022).

This new paradigm enables so-called *transfer learning*, which renders developing applications for financial use cases much more cost effective and accurate (see **Exhibit 1**). Pretrained models are widely available via open-source platforms, reducing model development to adapting the model's language understanding to its domain-specific uses. This process can be proprietary and requires only unlabeled data relevant to the target application, such as earnings call transcripts, financial news, or research reports, which are often abundant. Fine-tuning the adapted model to a specific task, such as sentiment analysis or return prediction, requires labeled data, which may need human expertise. Still, it can be relatively small, measured in thousands of documents rather than (hundreds of) millions.

As a result, it has become easier to unlock investment insights hidden in unstructured text data. Established applications such as sentiment analysis have evolved to become more sophisticated and effective, while new applications are beginning to emerge.

Applications and Use Cases along the Investment Process

NLP applications aim to extract relevant information from language data and are thus very versatile. As a result, *use cases vary widely, depending on how investors use text data* throughout the investment process and where NLP can add value. The investment strategy, asset classes, holding periods, and degree of process automation further shape the specifics of suitable applications. The availability of human, technical, and financial resources affects the sophistication of applications and influences decisions about external sourcing and in-house development.

The most popular applications cover a spectrum from data management, research support, and idea generation to alpha signal generation:

- On one end of the spectrum are applications that aim to boost the *productivity of research analysts* faced with growing amounts of potentially relevant text data. NLP solutions that focus on efficiency range from tagging, intelligent search, text summarization, and document comparison to the discovery of topics based on ML that helps organize large numbers of documents.
- On the other end are applications that extract *predictive signals to generate alpha* using state-of-the-art deep learning models. The outputs can inform discretionary and systematic portfolio management decisions or feed into further models to drive an automated strategy.

Along this spectrum exist numerous applications that *quantify or score text content* for specific characteristics deemed insightful and complementary to the content alone. Well-known examples include the analysis of sentiment, which can be defined in various ways, as we will see later. Alternative targets include ESG scores that measure how well corporate communication aligns with certain concepts. Approaches have tracked the evolution of NLP technology, as I will discuss in more detail.

Various inputs and techniques underlie most NLP applications. *Data sources* are the starting point and have multiplied and diversified. They range from public sources, such as machine-readable SEC filings, to familiar aggregators, such as Bloomberg, S&P Global, or Refinitiv (now LSEG Labs). Primary news or social media sources are now offering data to investors, as do companies that generate text data. Web-scraping produces potentially more exclusive data but is more resource intensive. Vendors have emerged that collect and aggregate text data, such as earnings calls. Depending on the source format, additional data preparation may be required to make raw data machine readable using, for example, image recognition, OCR (optical character recognition), or PDF (portable document format) extraction.

Text preprocessing segments raw text into tokens and higher-order elements, such as paragraphs, sentences, and phrases. Another critical step to enable further analysis is tagging text elements with linguistic annotations and other metadata, in particular, entity information, to resolve references to investment targets. Alternatively, data providers also provide "machine-ready" news that incorporates some metadata.

Investment firms that seriously pursue NLP tend to build proprietary pipelines using open-source tools that convert text data into custom model inputs because of the high added value at this step. Like other AI disciplines, *open-source software* has played an essential role in pushing the NLP research frontier, facilitating widespread technology adoption, and developing applications. A central reason is that the NLP workflow has common steps that can share tools. The Python language community has been particularly active in this regard. A popular Python library for these tasks is spaCy.

As a result, the development cost of custom NLP applications has dropped significantly and has become accessible to medium-sized firms. Furthermore, vendors have emerged that provide specialized solutions for certain workflow aspects and end-to-end solutions for specific applications: The estimated value of the global NLP market was USD10 billion in 2019 and is projected to grow to USD33 billion by 2025 (Lewington and Sartenaer 2021).

An excellent example of the evolution and broadening availability of NLP solutions is *Kensho*, which aims to make advanced information processing capabilities available to market participants beyond leading quant hedge funds. The company started in 2012 and was acquired by Standard & Poor's in 2018, setting a record for AI-driven deals at USD550 million, exceeding the price paid by Google for DeepMind. Kensho had developed an application named Warren (after Warren Buffett) with a simple textbox interface that offered investment advice to complex questions posed in plain English. It could access millions of data points, including drug approvals, economic reports, monetary policy changes, and political events, and could reveal their impact on nearly any traded financial asset worldwide. The company also provides an automated contextual analysis of changes in news flow, earnings reports, and analyst ratings. Early adopters included Goldman Sachs, which primarily used it to advise investment clients who contacted its trading desks.

Solutions have since evolved to focus on unstructured data and cover several techniques and applications mentioned previously. Namely, the company transcribes earnings calls, extracts data from PDF documents, detects and resolves named entities, such as companies, and links them to established identifiers. It also offers ML solutions, such as fine-tuning text classifiers and high-quality training sets, including 5,000 hours of earnings call recordings for multimodal modeling.

A *2020 survey of 20 customers* by the LSEG (London Stock Exchange Group) Labs (formerly Refinitiv) found that *65% had begun to use NLP applications*. Of this group, around one-third were just getting started, and another third were expending significant resources to identify use cases and deploy solutions. In contrast, the final third considered NLP fundamental and had already implemented the most advanced technologies, including deep learning models (Lewington and Sartenaer 2021).

An analysis of automated downloads of SEC filings further illustrates the *value different organizations place on NLP applications* (Cao et al. 2020). **Exhibit 2** displays the number of downloads over 2004–2017 by 13F filers. Among the most active consumers of machine-readable corporate filings are leading quantitative hedge funds, from Renaissance Technologies to D. E. Shaw. These institutions are the most likely to develop cutting-edge applications, such as the end-to-end training of models that combine text and other data sources to directly predict such outcomes of interest as returns for use in automated trading strategies. Institutions that heavily rely on research analysts, such as asset managers (but also quantitative hedge funds), certainly use productivity-enhancing NLP applications to efficiently process the large amounts of information contained in regulatory filings and other text documents.

I now proceed to describe applications that facilitate text data management to support the research process and lay the groundwork for more sophisticated applications. Then, I discuss how sentiment and other scoring tools have evolved with improved technology. Finally, I sketch applications on the horizon at the research frontier.

I illustrate these applications using publicly available datasets containing real-life financial news (see **Exhibit 3**). The financial phrasebank contains 4,840 sentences from English language news about companies trading on the Finnish stock exchange. Annotators have hand labeled each sentence as positive, neutral, or negative. I use the portion of the dataset where all annotators agree. The US financial news dataset consists of over 200,000 financial news articles on US companies sourced from various websites.

Getting Started: Organizing and Tagging Large Amounts of Text Data

The dramatic increase in text data volume generates new opportunities to detect a valuable signal but risks inundating the investment process with noise. To boost research analyst productivity and focus investment manager attention on critical events, tools that automate search, categorization, and the detection of relevant information are paramount.

The first steps toward this goal involve applying the initial stages of the NLP workflow described previously. More specifically, they include the following applications:

- **Named entity recognition (NER):** The automated tagging of objects of interest, such as corporations, people, or events

Exhibit 2. Downloads for Select 13F Filers, 2004–2017

Investment Firm	No. of Downloads	Type
Renaissance Technologies	536,753	Quantitative hedge fund
Two Sigma Investments	515,255	Quantitative hedge fund
Barclays Capital	377,280	Financial conglomerate with asset management
JPMorgan Chase	154,475	Financial conglomerate with asset management
Point72	104,337	Quantitative hedge fund
Wells Fargo	94,261	Financial conglomerate with asset management
Morgan Stanley	91,522	Investment bank with asset management
Citadel LLC	82,375	Quantitative hedge fund
RBC Capital Markets	79,469	Financial conglomerate with asset management
D. E. Shaw & Co.	67,838	Quantitative hedge fund
UBS AG	64,029	Financial conglomerate with asset management
Deutsche Bank AG	55,825	Investment bank with asset management
Union Bank of California	50,938	Full-service bank with private wealth management
Squarepoint	48,678	Quantitative hedge fund
Jefferies Group	47,926	Investment bank with asset management
Stifel, Nicolaus & Company	24,759	Investment bank with asset management
Piper Jaffray (now Piper Sandler)	18,604	Investment bank with asset management
Lazard	18,290	Investment bank with asset management
Oppenheimer & Co.	15,203	Investment bank with asset management
Northern Trust Corporation	11,916	Financial conglomerate with asset management

Exhibit 3. Reading the Financial Phrasebank

```
def read_phrase_bank():
     """Load sentences and annotations and return unique values"""
     label_to_score = {'negative': -1, 'neutral': 0, 'positive': 1}
     file_path = Path('financial_phrase_bank',
'Sentences_AllAgree.txt')
     text = file_path.read_text(encoding='latin1').split('\n')
     df = pd.DataFrame([s.split('@') for s in text],
                    columns=['sentence', 'sentiment'])
     df.sentiment = df.sentiment.map(label_to_score)
     return df.dropna().drop_duplicates()
```

- **Linguistic annotation:** Labeling of text elements regarding their grammatical function and other aspects
- **Topic modeling:** The identification of themes present in documents for automated categorization

I illustrate how each works and can, in turn, add value to the investment process.

Programmatic Tagging of Content with NER

Named entity recognition detects real-world concepts of interest, such as companies, individuals, locations, or custom entities in the text, allowing for spelling and other variations to ensure robustness across various sources. Among many different uses, it also supports compliance processes—for example, in anti-money laundering—because it can quickly flag critical items that require further diligence (sanction lists, politically exposed persons, etc.).

It automatically *annotates documents with tags or metadata* to enable the real-time detection of such concepts. Furthermore, it permits the integration of different datasets, such as news data that refer to a company by name and price data that identify relevant securities by ticker. It also lays the foundation for analyzing relationships between different concepts, such as competitors or members of a supply chain, or for comparing information for a given object at other points in time.

NER implementations can rely on predefined dictionaries that map words to concepts or learn these mappings from manually labeled data. The following code snippet (see **Exhibit 4**) shows some of the annotations that spaCy provides out of the box—that is, without custom training—for a given sentence: It recognizes nationality, company names, and ticker/exchange symbols, as well as the date and monetary values.

NER is a *foundational tool* that forms the basis for many other applications. Any investment firm that aims to automate the processing of text documents will require its results, whether it aims to use interactive applications to augment analysts or to build more sophisticated ML models that use documents.

Linguistic Annotation: Parsing Dependencies among Sentence Elements

While NER detects items in isolation, linguistic annotations establish *functional relationships among text elements* in a sentence. This technique reverse engineers the connections between objects and actions using the grammatical structure and provides vital inputs into downstream efforts at recovering the meaning of the content.

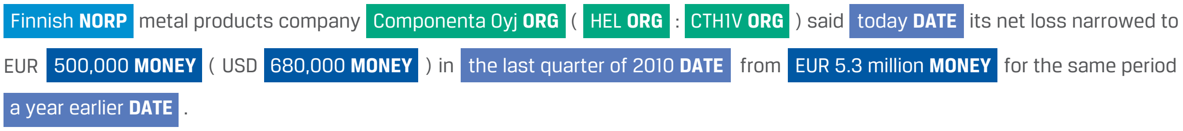
The details are beyond the scope of this chapter, but the code snippet in **Exhibit 5** illustrates how spaCy annotates a fragment of the sample sentence from Exhibit 4.

It simultaneously identifies the grammatical roles of words, such as nouns, verbs, and prepositions; the dependencies among them (e.g., "company" is further defined by "Finnish metal products"); and coherent phrases.

Downstream applications can use these metadata and the resulting dependency parsing tree to screen sentences for specific patterns and phrases, such as "net loss narrowed," with more positive implications than the individual words in isolation might suggest (Marinov 2019). An example is custom sentiment analysis that encodes specific observations about meaningful statements.

Exhibit 4. Named Entity Recognition with spaCy

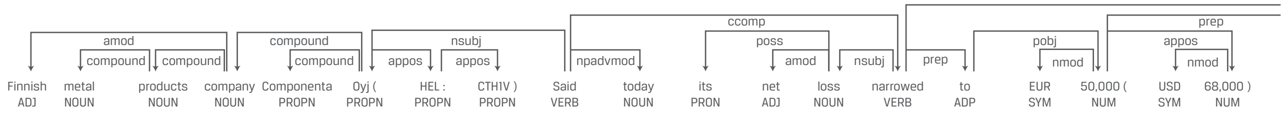
```
import spacy
from spacy import displacy
nlp = spacy.load('en_core_web_trf')
phrases = read_phrase_bank()
sentences = phrases.sentence.tolist()
doc = nlp(sentences[331])
displacy.render(doc, style='ent', jupyter=True, options={'compact':
True})
```



Finnish **NORP** metal products company Componenta Oyj **ORG** (HEL **ORG** : CTH1V **ORG**) said today **DATE** its net loss narrowed to EUR 500,000 **MONEY** (USD 680,000 **MONEY**) in the last quarter of 2010 **DATE** from EUR 5.3 million **MONEY** for the same period a year earlier **DATE** .

Exhibit 5. Grammatical Dependency Relations with spaCy

```
doc = nlp(sents[331])
displacy.render(doc, style='dep', jupyter=True, options={'compact':
True, 'distance': 100})
```



To leverage the potential value of linguistic annotations, organizations need a dedicated data science team with NLP expertise to further process the results. Hence, it would rarely be used outside larger investment organizations.

Automatic Discovery of Themes across Documents with Topic Modeling

A topic model aims to *discover hidden themes* across a body of documents and computes the relevance of these themes to each of them. The most popular technique is latent Dirichlet allocation, or LDA (Blei, Ng, and Jordan 2003). In this context, a theme consists of terms that are more likely to appear jointly in documents that represent this theme than elsewhere. The discovery process is fully automated given a target number of topics, which can be optimized using statistical goodness-of-fit measures.

The results share a drawback common to *unsupervised learning*: The lack of a ground truth benchmark implies the absence of objective performance metrics. In other words, the value of a particular set of topics identified by the model depends on how useful they are to the downstream application. Moreover, topics do not come with handy labels; each one corresponds to a list of words with associated weights (but humans tend to quickly grasp the concepts behind a topic upon inspection). These practical challenges notwithstanding, topic models are among the most effective tools for organizing large amounts of text in a meaningful and automatic way.

Investment firms use topic models as tools that *facilitate the research process*. They are useful because they summarize large amounts of text in a much more manageable number of themes that are easy to track over time. At the same time, the content of individual documents transparently relates to these topics. Analysts can use this information to compare how the relevance of topics changes over time or across related companies and quickly identify the relevant text passages. Implementations used to require custom developments but can now rely to a much greater degree on third-party APIs. As a result, this technology is becoming increasingly accessible to small and medium-sized firms.

Bryan Kelly, of AQR and Yale University, and co-authors illustrate a practical use of LDA by identifying 180 topics in over 800,000 *Wall Street Journal* articles covering 1984–2017 (Bybee, Kelly, Manela, and Xiu 2021). The authors maintain an interactive website (www.structureofnews.com) that illustrates how textual analysis can summarize the state of the economy. The uncovered *structure of economic news* shows several benefits of the topic model output:

- The documents most relevant for a given topic are easy to retrieve at any point in time because each text has a topic weight.
- The relative prevalence of themes also matters because it reflects the attention that news pays to them over time. In other words, the mix of topics at any time reflects the current state of the economy and does correlate with macroeconomic indicators. Moreover, it explains 25% of aggregate market returns when used as an input to a statistical model and predicts economic events, such as recessions, beyond standard indicators.

Two Sigma illustrates how LDA can add value when applied to *Federal Open Market Committee meeting minutes* to quantify the relative weight of various economic issues over time. The company's analysis shows how financial market topics gained relative weight, at the expense of growth-related themes, at the beginning of Alan Greenspan's tenure, underpinning the market's perception of the Greenspan put at the time (Saret and Mitra 2016).

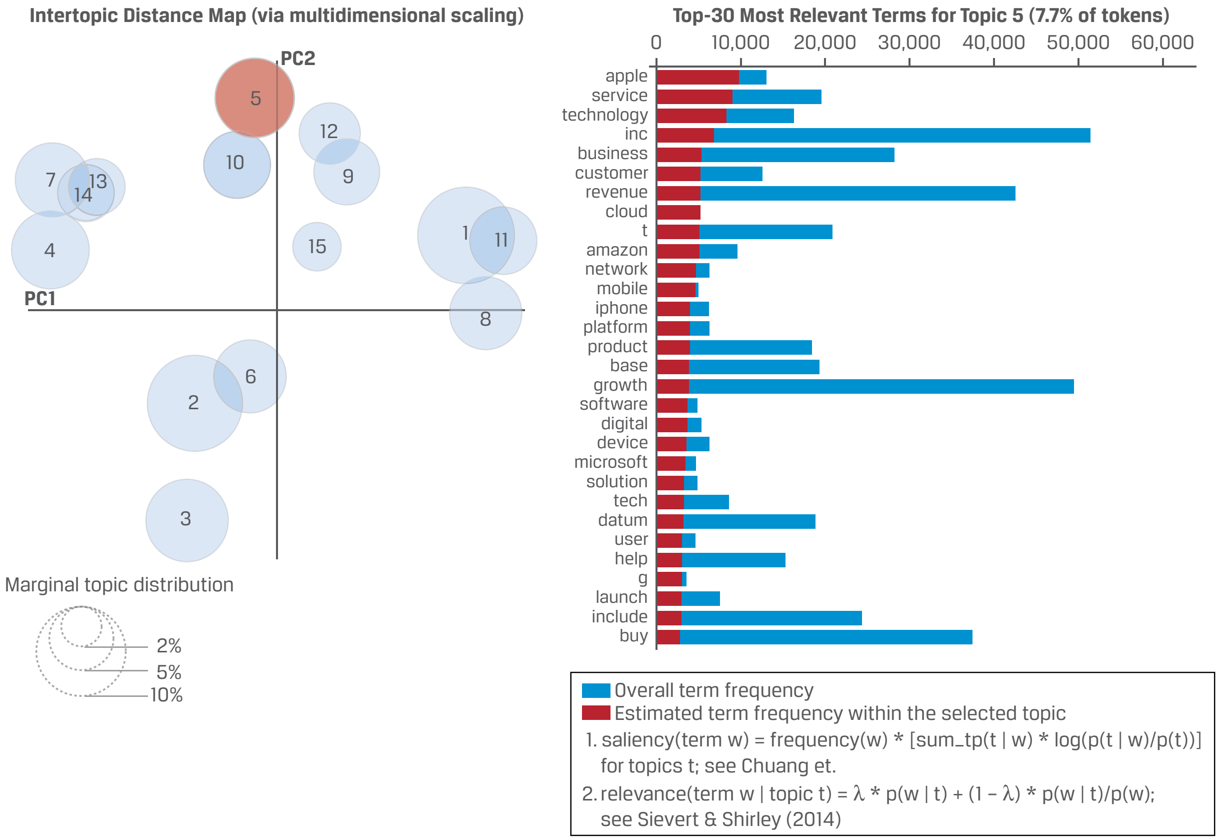
Exhibit 6 illustrates the selection of a random sample of 25,000 financial news articles with 250–750 words each, the preprocessing of the text with spaCy to remove less informative terms, and the use of Gensim's LDA implementation to identify 15 topics.

We then use pyLDAvis to visualize the relative importance of the 15 topics interactively. **Exhibit 7** displays the vocabulary frequency for each theme compared to its frequency across all documents. Topic 5, for instance, discusses technology companies.

Exhibit 6. Latent Dirichlet Allocation Using Gensim

```
df = pd.read_csv('us_news.csv', usecols=['content']).drop_duplicates()
df = df[df.str.split().str.len().between(250, 750)].sample(n=25000)

nlp = spacy.load('en_core_web_md')
preprocessed_docs = []
to_remove = ['ADV', 'PRON', 'CCONJ', 'PUNCT', 'PART', 'DET', 'ADP',
'SPACE', 'NUM', 'SYM']
for doc in tqdm(nlp.pipe(df.tolist(), n_process=-1)):
     preprocessed_docs.append([t.lemma_.lower() for t in doc
                           if t.pos_ not in to_remove and not
t.is_stop and t.is_alpha])
len(nlp.vocab) # 117,319 tokens
dictionary = Dictionary(preprocessed_docs)
dictionary.filter_extremes(no_below=5, no_above=0.2, keep_n=5000)
LdaMulticore(corpus=corpus, id2word=dictionary, iterations=50,

     num_topics=num_topics, workers=4, passes=25, random_state=100)
```

Exhibit 7. Topic Visualization Using PyLDAvis



Sources: PyLDAvis using information from Chuang, Manning, and Heer (2012) and Sievert and Shirley (2014).

Scoring Text for Sentiment Analysis

The impact of news encoded in text data on prices has long motivated investors to design summary metrics that could anticipate the market's reaction to news content. Sentiment analysis is by far the most popular NLP application: Systematic attempts to score market sentiment regarding new information on an asset and to invest accordingly have been documented for almost 100 years (Cowles 1933). Computerized sentiment analysis at scale gained momentum after 2000.

Sentiment analysis is very popular because scores for news stories or corporate communications are available from various sources, including the ubiquitous Bloomberg terminal. Research analysts can use such scores to quantify and compare trends across a coverage universe, and portfolio managers can evaluate news flow quality "at a glimpse." Scores can also be used as trading signals (and are often promoted to this end).

However, there are many ways to derive such scores, as I will discuss later, not least because the term "sentiment" offers room for interpretation. Custom implementations tend to provide more valuable information because they control the definition of the outcome. However, they rely on supervised learning, which, in turn, requires costly human expertise to label data. It also requires in-house data science capabilities or trusted vendors.

Measuring Text Characteristics by Scoring Text Elements

Initially, practitioners relied on word lists developed in other disciplines, such as psychology, to score the positive or negative sentiment likely expressed by individual words. To evaluate the polarity of an entire document, they aggregated the sentiment scores.

In 2011, Loughran and McDonald demonstrated how the specific use of language in finance requires tailored approaches to avoid misinterpretation. For instance, investors tend to focus more on negative aspects. However, they face the challenge that corporate communication carefully controls positive and negative word frequencies in earnings calls and other disclosures (Loughran and McDonald 2011).

Beyond Sentiment: Capturing Intentions and Hidden Meaning

Loughran and McDonald (2011) introduced word lists specific to finance to capture both tone and behavioral aspects that investors care about. Besides positive or negative language, investment managers pay attention to the degree of confidence surrounding forward-looking statements to gain insights into the attitudes of more informed insiders.

Common word lists thus emphasize hints at obfuscation or deception (such as the overuse of favorable terms) or uncertainty (for example, via the possibly unintentional use of such "weak modal" words as "may" or "could," in earnings calls). Therefore, Loughran and McDonald added lists that flagged when the word choice strengthens or weakens material statements. **Exhibit 8** lists the five most common terms for the original seven lists, which have received minor updates since publication.

Loughran and McDonald further differentiated these lists by channel of communication to address additional bias. They also accounted for differences in the meaning of words in, for instance, regulatory filings, where the term "auditor" does not have the negative connotation typically perceived in business news.

Another text characteristic that investors frequently scrutinize is the readability of text, measured by such metrics as the number of words per sentence or the number of syllables per word. The rationale is that overly complex language may indicate an intent to mislead or distract the reader. Just as with sentiment analysis, the definition of

Exhibit 8. Most Common Terms in Select Word Lists

Negative	Positive	Uncertainty	Litigious	Strong_Modal	Weak_Modal	Constraining
LOSS	GAIN	MAY	SHALL	WILL	MAY	REQUIRED
LOSSES	GAINS	COULD	AMENDED	MUST	COULD	OBLIGATIONS
TERMINATION	ABLE	APPROXIMATELY	CONTRACTS	BEST	POSSIBLE	REQUIREMENTS
AGAINST	GOOD	RISK	HEREIN	HIGHEST	MIGHT	RESTRICTED
IMPAIRMENT	ADVANCES	RISKS	LAW	NEVER	DEPEND	IMPAIRMENT

Source: Loughran and McDonald (2011).

readability requires sensitivity to the context: The language in a financial report has a different baseline complexity than financial news, and the criteria and thresholds need to be adapted accordingly. Moreover, critics maintain that the complexity of language primarily reflects the size of an organization (Loughran and McDonald 2020).

From Sentiment Scores to Investment Decisions

Subsequently, vendors emerged that offered domain-specific word lists and ready-made sentiment scores. Document scores often become inputs to analysts' research process to compare their own values across a universe of competitors. Analysts also combine sentiment scores with the results of a topic model to quantify how differently alternative investment targets discuss similar themes.

Or, sentiment scores can feed into a model that aims to explain or predict returns. Academic evaluations of this approach tend to find a statistically significant yet economically small association between document sentiment and subsequent returns.

However, measuring text characteristics with dictionaries assembled by domain experts has several weaknesses:

1. Interpreting text by evaluating words in isolation ignores a lot of information.
2. Word lists are backward looking: Even if the domain expertise that maps terms to polarity scores successfully captures signal rather than noise, this relationship may break down at any time.
3. Moreover, if the communicators are aware of the word lists used by investors, they may adjust their word choice to influence the market favorably.

Word Lists in Practice

The Loughran–McDonald word lists are in the public domain, and the code example in **Exhibit 9** illustrates their use and predictive accuracy on the hand-labeled phrasebank dataset.

The traditional Loughran–McDonald word list approach achieves an accuracy of 66.03%. A defining characteristic of the methods described in this section is that the scoring or classification rules are typically handcrafted rather than learned by a machine from data. We now turn to a data-driven alternative.

Modern NLP: Learning to Predict Arbitrary Outcomes from Text Data

The evolution of sentiment analysis tracks the progress in NLP from rule-based to data-driven approaches outlined earlier in this chapter. The rapid adoption of state-of-the-art modeling techniques based on transfer learning are making sophisticated applications much more accessible. As in other industries, custom models that predict targets of interest are becoming commonplace beyond industry leaders.

These applications offer enormous flexibility because input data and target outcomes can be designed if training data are available. *Dataset curation* often poses the costliest hurdle: It requires the annotation of the input documents and other data with custom labels that reflect the historical realizations of the target outcome. Such outcomes may be price movements over a given horizon. However, they can also be more subtle, such as the appropriate confidence in management forecasts made during an earnings call, the adherence to some ESG standard, or compliance with specific regulations. If the targets are subsequent returns,

A Caveat: Campbell's Law

The last point in the preceding list illustrates a broader caveat to the analysis of human language: Humans adapt and modify their behavior to game the system. Campbell's law, also cited as Goodhart's law, originally formulated this phenomenon. Marylin Strathern (1997, p. 308) summarized it as follows: "When a measure becomes a target, it ceases to be a good measure."

Careful analysis of corporate communication after the release of the Loughran–McDonald word lists uncovers significant changes in word choice, particularly by companies most likely subjected to automated sentiment scoring (Cao et al. 2020).

Given the reality of executives' adaptation to automated sentiment analysis, it can be more effective to focus on how analysts ask questions because they do not have equally obvious incentives to pretend (Mannix 2022).

Exhibit 9. Sentiment Analysis Using Loughran–McDonald Word Lists

```
def download_lm():
   """Source latest LM word-score dictionary"""
   url = 'https://drive.google.com/file/d/17CmUZM9hGUdGYjCXcjQLyybjTrcjrhik/view?usp=sharing'
   url = 'https://drive.google.com/uc?id=' + url.split('/')[-2]
   return pd.read_csv(url)

def lm_counter_by_theme():
   """Return list of terms associated with the seven LM themes"""
   df = download_lm()
   df = df.rename(columns=str.lower).assign(word=lambda x: x.word.str.lower())
   return {theme: Counter(df.loc[df[theme].gt(0), 'word'].tolist()) for theme in lm_sentiment}

def get_lm_sentiment(df):
   lm = lm_counter_by_theme()
   phrases['sentence_'] = phrases.sentence.str.lower().str.split().apply(Counter)
   for direction in ['positive', 'negative']:
       df[direction] = (df.sentence_.apply(lambda x: len(list((x &
lm[direction]).elements())))
                        .clip(lower=-1, upper=1))

   df = df.drop(['sentence_'], axis=1)
   df['lm'] = df.positive.sub(df.negative)
   df['delta'] = df.sentiment.sub(df.lm)
   return df.drop(['positive', 'negative'], axis=1)

labels = ['negative', 'neutral', 'positive']
phrases = read_phrase_bank()
phrases = get_lm_sentiment(phrases)

lm_cm = confusion_matrix(y_true=phrases.sentiment, y_pred=phrases.lm)
lm_cm = pd.DataFrame(lm_cm, columns=labels, index=labels)
lm_acc = accuracy_score(y_true=phrases.sentiment, y_pred=phrases.lm)

       negative  neutral  positive
negative  120        177     6
neutral   68         1284    39
positive  70         409     91
```

one can automate the annotation process. However, if they depend on domain expertise, labeling can turn into a time-intensive use of human capital.

Numerous vendors offer annotation services because this step is critical for state-of-the-art NLP projects across industries. These include solutions that themselves rely on AI, such as Scale AI, Snorkel AI, and Labelbox, startups valued over USD1 billion each.[1] There are also efficiency-enhancing software tools, such as Prodigy.[2]

Currently, NLP applications that rely on custom models based on transfer learning are beginning to proliferate outside the largest quantitative hedge funds with dedicated research teams. Proof-of-concept developments can today be put together by smaller teams so that R&D pilots are more common in mid-sized investment firms.

From Rule-Based to Data-Driven Approaches

Over the past few years, substantial advances in deep learning have moved the frontier for state-of-the-art applications from rule-based to data-driven approaches. Here, machines learn directly from data how to interpret

[1] For more information, go to the companies' respective websites: https://scale.com/; https://snorkel.ai/; https://labelbox.com/.

[2] Go to https://prodi.gy/.

language and predict sentiment scores or other outcomes of interest, such as subsequent asset returns or ESG scores, more accurately. In other words, cutting-edge applications eschew handcrafted lists and rely on ML models to identify patterns relevant to the predictive target.

Approaches range from white-box models that learn transparent word lists from data (Ke, Kelly, and Xiu 2019) to large language models that are pretrained on the generic text and fine-tuned to domain-specific documents (Araci 2019; Huang, Wang, and Yang, forthcoming).

As with a rule-based approach, a downstream model that aims to predict returns can combine sentiment scores produced by a language model with other inputs. Alternatively, models can be designed and trained in an end-to-end fashion.

The benefits include exploiting much richer semantic input than word lists and the flexibility to train a model to learn outcomes directly relevant to the investment strategy. However, the custom approach requires curated, labeled data: Each training document requires annotation with the corresponding ground truth outcome.

Risks include overfitting to noise in the training sample and limited ability to transfer the model's learning from the pretraining inputs to the domain-specific documents. Fortunately, these risks can be managed and measured before using the resulting model in a production setting.

Modern NLP and Transfer Learning in Practice with Hugging Face and FinBERT-Tone

Hugging Face is an NLP startup based in New York City with USD160 million in funding and a USD2 billion valuation that offers state-of-the-art custom solutions. It also hosts many pretrained and fine-tuned language models and allows for free downloads.

One of the models is FinBERT, a large-scale BERT model based on the transformer architecture with almost 1 million downloads per month. It trained on SEC reports, earnings call transcripts, and analyst reports containing close to 5 billion tokens, with further fine-tuning on the hand-labeled phrasebank documents.

The code in **Exhibit 10** illustrates how to access the fine-tuned model and evaluate its predictive performance relative to the Loughran–McDonald word lists.[3]

FinBERT achieves 91.7% accuracy, a 39% improvement over the Loughran–McDonald approach.

The sample sentence we used for Exhibit 10 to illustrate basic NLP techniques shows why the language model outperforms: "Finnish metal products company Componenta Oyj (HEL: CTH1V) said today its net loss narrowed to EUR500,000 (USD680,000) in the last quarter of 2010 from EUR 5.3 million for the same period a year earlier."

The phrase "net loss narrowed" requires taking at least some context into account, which simple word lists cannot do.

From the Research Frontier: Multimodal Models Integrate Text and Tone

Speech recognition is key to the analysis of companies' periodic earnings calls. Given the tight controls of the textual content and adaptation to popular sentiment scoring approaches, it can be informative to incorporate the speaker's style and tone into the analysis.

Recent so-called multimodal learning models jointly process different modalities, such as text and speech or video data, and learn how to interpret these complementary signals considering the target outcome. Such applications will likely become more common as more appropriately labeled data become available in house or from vendors.

Early results already suggest these applications can reveal what analysts ask about and how they do so, including which tone they use (and how company representatives respond). A quantified profile of the speakers' behavior is a valuable complement to textual analysis of the actual content and can be used to inform either human research or downstream models.

Summary and Conclusion

NLP applications have become much more widespread over the past 5–10 years because the greater amount of data increased demand and improved performance. At the same time, high-quality open-source software has become available and popular, reducing the cost of customized applications, developed either in house or through third-party vendors.

Many investment firms that rely on research driven by textual input use some form of automated text processing or scoring to manage the data explosion and human productivity while potentially improving results through the efficient discovery of hidden insights.

Applications that quantify text content regarding relevant metrics, such as sentiment or ESG scores, have become

[3] The comparison is not entirely fair because FinBERT has been trained on these documents. However, the comparison demonstrates how a language model overcomes the word list limitations mentioned previously.

Exhibit 10. Sentiment Analysis Using a Fine-Tuned Transformer Model

```
from transformers import BertTokenizer, BertForSequenceClassification
def get_finbert_sentiment(df):
   finbert = BertForSequenceClassification.from_pretrained('yiyanghkust/finbert-tone',
num_labels=3)
   tokenizer = BertTokenizer.from_pretrained('yiyanghkust/finbert-tone')

   sentiment = []
   for sents in chunks(df.sentence.tolist(), n=50):
       inputs = tokenizer(sents, return_tensors="pt", padding=True)
       outputs = finbert(**inputs)[0]
       sentiment.extend(list(np.argmax(outputs.detach().numpy(), axis=1)))

   df['finbert'] = sentiment
   df.finbert = df.finbert.apply(lambda x: -1 if x == 2 else x)
   return df

phrases = get_finbert_sentiment(phrases)

fin_cm = confusion_matrix(y_true=phrases.sentiment,
                         y_pred=phrases.finbert)
       negative  neutral  positive
negative  120         177     6
neutral   68          1284    39
positive  70          409     91
```

very common. Most importantly, the unprecedented breakthroughs in deep learning for language and transfer learning have boosted the accuracy of such measures, especially when tailored to the specific use case relevant to an investor.

Progress will almost certainly continue in the direction of large language models that learn to understand domain-specific language well in cost-effective ways. The challenges will consist of harnessing these new technologies to automate existing processes and identifying predictive targets that add value to the investment process.

References

Araci, D. 2019. "FinBERT: Financial Sentiment Analysis with Pre-Trained Language Models." Cornell University, arXiv:1908.10063 (27 August).

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

Bommasani, R., Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. "On the Opportunities and Risks of Foundation Models." Cornell University, arXiv:2108.07258 (16 August).

Bybee, L., B. T. Kelly, A. Manela, and D. Xiu. 2021. "Business News and Business Cycles." NBER Working Paper 29344 (October).

Cao, S., W. Jiang, B. Yang, and A. L. Zhang. 2020. "How to Talk When a Machine Is Listening?: Corporate Disclosure in the Age of AI." NBER Working Paper 27950 (October).

Cowles, A. 1933. "Can Stock Market Forecasters Forecast?" *Econometrica* 1 (3): 309–24.

Chuang, J., Christopher D. Manning, and Jeffrey Heer. 2012. "Termite: Visualization Techniques for Assessing Textual Topic Models." In *International Working Conference on Advanced Visual Interfaces (AVI)*, 74–77.

Firth, J. R. 1957. "Applications of General Linguistics." *Transactions of the Philological Society* 56 (1): 1–14.

Harbert, T. 2021. "Tapping the Power of Unstructured Data." MIT Sloan (1 February). https://mitsloan.mit.edu/ideas-made-to-matter/tapping-power-unstructured-data.

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

Ke, Z. T., B. T. Kelly, and D. Xiu. 2019. "Predicting Returns with Text Data." NBER Working Paper 26186 (August).

Lewington, D., and L. Sartenaer. 2021. "NLP in Financial Services." LSEG Labs.

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

Loughran, T., and B. McDonald. 2020. "Textual Analysis in Finance." *Annual Review of Financial Economics* 12: 357–75.

Mannix, R. 2022. "Analysts Reveal More than Company Execs on Earnings Calls—AB." Risk.net (23 February). www.risk.net/node/7932076.

Marinov, S. 2019. "Natural Language Processing In Finance: Shakespeare without the Monkeys." Man Institute (July).

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. "Distributed Representations of Words and Phrases and Their Compositionality." *Proceedings of the 26th International Conference on Neural Information Processing Systems* 2: 3111–19.

Pollack, A. 1983. "Technology; The Computer as Translator." *New York Times* (28 April). www.nytimes.com/1983/04/28/business/technology-the-computer-as-translator.html.

Polu, S., and I. Sutskever. 2020. "Generative Language Modeling for Automated Theorem Proving." Cornell University, arXiv:2009.03393 (7 September).

Popel, M., M. Tomkova, J. Tomek, Ł. Kaiser, J. Uszkoreit, O. Bojar, and Z. Žabokrtský. 2020. "Transforming Machine Translation: A Deep Learning System Reaches News Translation Quality Comparable to Human Professionals." *Nature Communications* 11.

Saret, J. N., and S. Mitra. 2016. "An AI Approach to Fed Watching." Two Sigma. www.twosigma.com/articles/an-ai-approach-to-fed-watching/.

Sievert, Carson, and Kenneth Shirley. 2014. "LDAvis: A Method for Visualizing and Interpreting Topics." In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70. Baltimore: Association for Computational Linguistics.

Srivastava, A., A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al. 2022. "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models." Cornell University, arXiv:2206.04615 (10 June).

Strathern, Marilyn. 1997. "'Improving Ratings': Audit in the British University System." *European Review* 5 (3): 305–21.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. 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.

Wang, A., Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. 2019. "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems." Cornell University, arXiv:1905.00537 (2 May).

6. EXTRACTING TEXT-BASED ESG INSIGHTS: A HANDS-ON GUIDE

Tal Sansani, CFA

Founder, Off-Script Systems, and Co-Founder, CultureLine.ai

Mikhail Samonov, CFA

Founder, Two Centuries Investments, and Co-Founder, CultureLine.ai

Introduction

This chapter presents several increasingly prevalent and effective applications of natural language processing (NLP) techniques in environmental, social, and governance (ESG) investing. Our guide simultaneously progresses down two tracks: the recent history of quantum leaps in NLP modeling and applying those advances to the present-day, high-stakes challenges of ESG investing. Modern NLP thrives in a vibrant, open-source environment, allowing investment practitioners to stand on the shoulders of AI giants by leveraging immense cumulative knowledge and computational processing power toward customized investment solutions. The key features of NLP—seamless customization and dynamic adaptability—are uniquely suited for the rapidly evolving language and standards in the ESG ecosystem.

With approximately USD17 trillion in equity capital invested in ESG-oriented funds (US SIF Foundation 2020), NLP-based insights and signals already underpin a variety of ESG datasets, widely used ratings and scores, alternative data offerings, and proprietary buy-side investment tools. These relatively new tools are helping address many of the most-cited challenges faced by institutional ESG investors—a lack of standardized data from third-party providers, limited disclosures from companies, and subjectivity of metrics (Cerulli Associates 2022).

This chapter covers a selection of foundational and state-of-the-art NLP advances, from word embeddings (word2vec, originating in 2013) to large language models (BERT, originating in 2018) to machine inference (zero-shot classifiers, originating in 2020). To support a learning journey, the text introduces NLP techniques in progressive layers, applying models that detect and connect *who* is mentioned, *what* topic is discussed, and *how* that topic is described in terms of tone and sentiment.

The simultaneous paths of this chapter are "hands-on" experiences with numerous examples, plain English explanations, and illustrations of how advanced NLP techniques can help identify ESG themes at the broad market level, the sector level, the individual company level, and even in the context of the more nuanced topic of "greenwashing." By showcasing recent NLP innovations, this chapter seeks to empower and ultimately inspire readers to apply NLP solutions to ESG problems.

ESG Investing: The Need for Adaptive Tools

As the concept of ESG has become part of investment vernacular (IFC 2004), the volume of data associated with related topics has grown exponentially. An entire subindustry of rating providers, index creators, gatekeepers, and industry coalitions has emerged to make sense of these growing inputs. Consequently, regulatory scrutiny has intensified, leading to the formation of various rules, guidelines, and proposals.

However, the dynamic nature of ESG metrics continues to pose challenges to the formation of industry standards. In just the last two years, investors have endured supply chain risks from COVID-19, diversity and inclusion initiatives resulting from racial justice protests, energy security concerns in the wake of the war in Ukraine, and climate risk related to carbon emissions. The fast pace and market volatility of these major global events demonstrate the need for flexible and adaptive tools for analysis and risk management. This chapter highlights how NLP can dynamically and objectively distill this complex landscape at scale.

ESG covers a multitude of valuation-relevant, intangibles-related topics missing from financial statements, such as product quality, supply chain resilience, human capital, and management quality, as well as other metrics traditionally associated with fundamental investment frameworks. The breadth of topics makes standardization challenging, which provides an opportunity for active investment research. ESG investing is yet another example of how market participants can gain an investment edge with novel and dynamic information that goes beyond traditional metrics.

NLP: Applying Human Insights at Scale

Natural language processing, which seeks to create statistical models that understand text and spoken words, is a rapidly growing branch of artificial intelligence (AI). In a short time, NLP models have gone from counting words in text to accurately inferring tone and meaning, a job that has historically been reserved for analysts. In a world awash in digital content, deciphering the trustworthiness, quality, context, and biases in language requires the collaboration of machines.

The nexus of ESG and NLP creates a scalable and customized way to extract timely insights from volumes of unstructured datasets, such as news (broad, curated, and widely consumed information), company and product review sites (crowdsourced information), company earnings calls, corporate social responsibility (CSR) reports, industry journals (information from practitioners and experts), and countless other sources. Whereas Deep Blue seeks to best humans at chess (machine over human), the most useful applications of AI in investment management put people and machines to work collaboratively, emphasizing each other's strengths in the hope of the sum being greater than the parts. In the investment context, ESG and fundamental analysts can apply their knowledge with NLP tools by creatively collaborating with in-house technical experts or consultants—teamwork that requires stepping out of traditional silos.

In the sections that follow, we demonstrate how NLP solutions are applied to some of the pressing questions surrounding ESG:

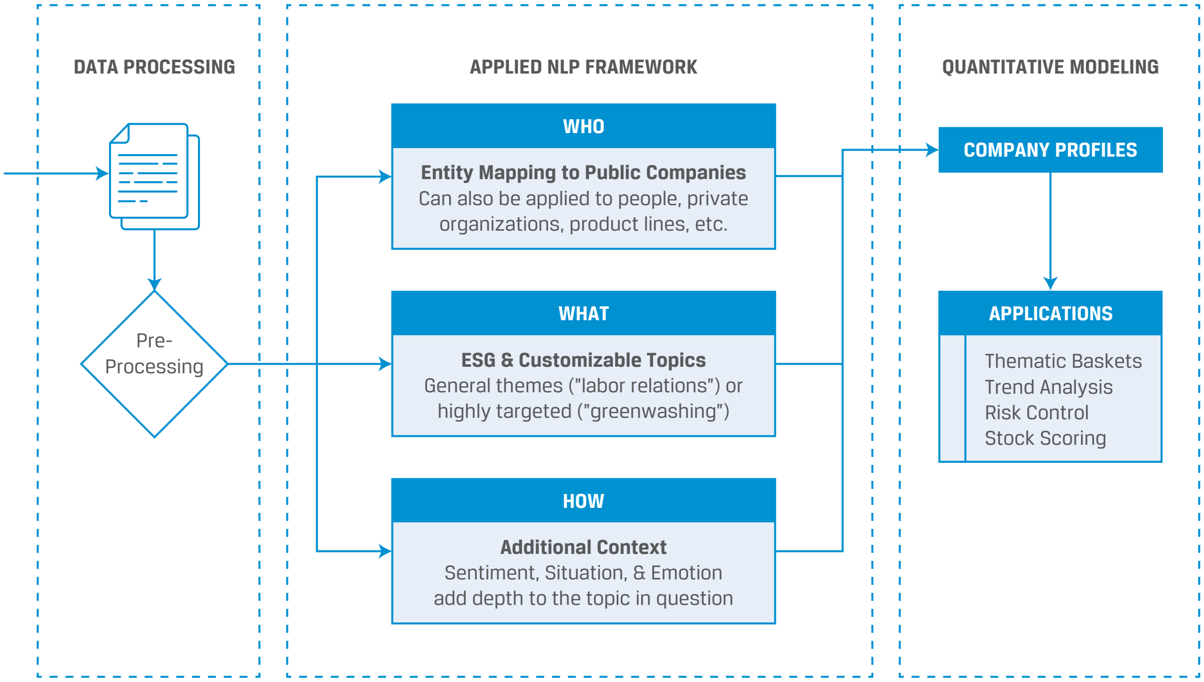
- Which issues matter most on a broad market level?
- What ESG issues are most prominent for a given sector or a particular company?
- Is the sentiment around a company's performance on sustainability variables worsening or improving?
- Is management spinning a story that is met with skepticism in the financial press?
- Which companies might have rising reputational and regulatory risk regarding "greenwashing"?

Extracting Text-Based ESG Insights

This chapter highlights ways to explore new ESG research or prototype new products by stacking relatively contemporary approaches on top of foundational NLP techniques.

These techniques can be learned and developed by layering on each technological advancement. To illustrate this progression, the framework in **Exhibit 1** combines three core approaches to extracting insights from text along with a guide to the initial structuring and sourcing of the information: *who* the text is about (publicly traded company), *what* the text is about (ESG topics), and *how* the topic is described (tone and sentiment).

Exhibit 1. Process Diagram: The Who, the What, and the How of NLP


DATA PROCESSING
Pre-Processing
APPLIED NLP FRAMEWORK
WHO
Entity Mapping to Public Companies
Can also be applied to people, private organizations, product lines, etc.
WHAT
ESG & Customizable Topics
General themes ("labor relations") or highly targeted ("greenwashing")
HOW
Additional Context
Sentiment, Situation, & Emotion add depth to the topic in question
QUANTITATIVE MODELING
COMPANY PROFILES
APPLICATIONS
Thematic Baskets
Trend Analysis
Risk Control
Stock Scoring

Corpus Creation: Financial News and Earnings Calls

Machine-driven insight is only as good as its data inputs. For this study, two complementary yet contrasting data sources—quarterly earnings transcripts and financial news—cover company developments across a variety of ESG issues from significantly different perspectives. The cost and time involved with fetching, cleaning, and organizing unstructured data typically represent a large first step; however, strong demand for text-based information sets (Henry and Krishna 2021) has led many data vendors to provide structured feeds of news[4] and earnings call transcripts[5] for NLP modeling, backtesting, and production processes.

Data evaluation

Preparing and aggregating data for analytical purposes are contextual and informed by how biases in the data could affect the model's output. In this study, earnings calls and financial news have three primary contrasting features: *timeliness* (calls are quarterly; news flow is periodic), *distribution of text* (earnings calls are done for virtually all publicly traded companies, while news is heavily skewed toward large companies), and *author's perspective* (calls are primary-source, first-person information originating from company executives, while news is third-party information).

Looking at the distribution and frequency of information flow in **Exhibit 2**, financial news is heavily biased toward large companies and coverage rises around earnings season. To correct for the large-company bias, metrics are aggregated at the company level and equally weighted to determine sector- and market-level trends (as opposed to weighting them by market capitalization or information share). For example, the thousands of articles discussing ESG issues important to Tesla are all mapped to Tesla, whereas the small number of articles covering ESG issues for Newmont Corporation are mapped to the mining company. When those companies are rolled up into one group for analysis, their ESG issues roll up with equal emphasis, meaning the ESG trends revealed will not be biased toward the largest companies that generate the most news.

Organizing the data: Entity mapping (the who)

Entity mapping is the process of connecting fragments of text (from sentences to full documents) to various entities (people, places, organizations, events, or specific product lines). Entity mapping answers the first important question: Who is this document or sentence about?

Mapping thousands of articles from multiple news outlets and earnings calls back to their entity (who) across thousands of companies is a challenging task, but the integrity of entity mapping drives the trustworthiness of subsequent insights. For example, references to "Meta," "Facebook," and "FB" need to map back to one entity, even though the references have changed over time. Other important questions arise in this process: Was the company mentioned in passing, or was it the focus of an in-depth, "long-read" article about the company? In addition to these details, online news outlets have different layouts, formats, advertisements, pop-ups, and other obstacles to systematically structuring text data for reliable downstream analysis.

Leveraging third-party technology

Although entity mapping is a daunting technical undertaking, sourcing and structuring vast quantities of new information from the internet has never been easier. For example, financial data vendors offer a wide range of off-the-shelf solutions, including structured, entity-mapped news feeds, so AI-driven insights are unencumbered by the difficulties of gathering and organizing data.

The example in **Exhibit 3** illustrates the simplicity, effectiveness, and scale of modern third-party solutions. Diffbot's query language (DQL) allows its users to query the web like any other database.

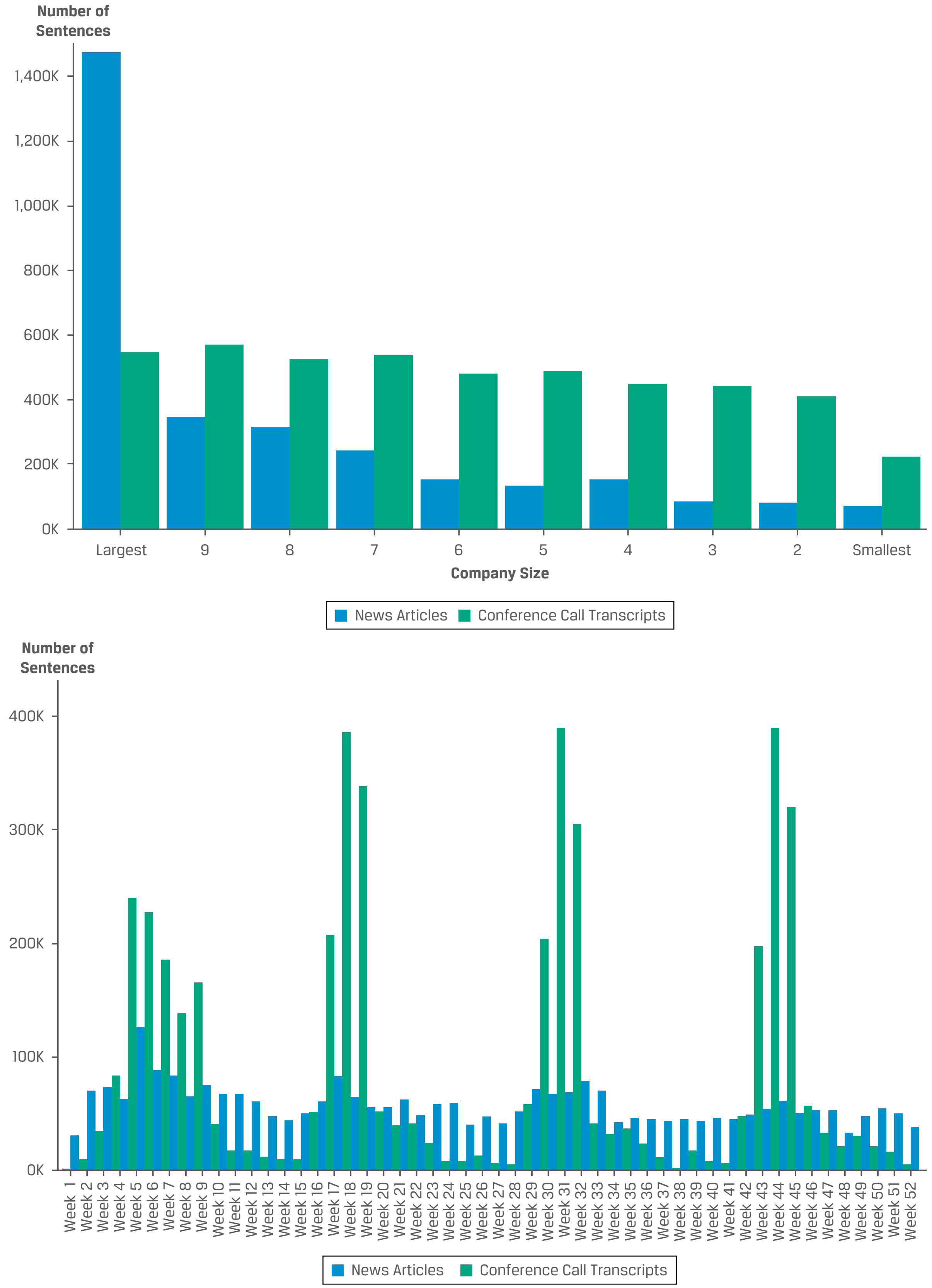
Five simple lines of code generate over 1,000 news articles in machine-readable form (e.g., CSV or JSON), each containing the article's text and title and the original article URL, augmented with rich metadata, such as the article's sentiment and the people, places, and general topics being discussed. While this example illustrates a single query, this paper draws on thousands of such queries, codified to collect English news articles for companies in the Russell 1000 Index, across 13 of the most popular financial news sources.[6] Modern, NLP-driven application programming

[4] Over 30 news analytics vendors are listed at the AlternativeData.org database of data providers: https://alternativedata.org/data-providers//search,news.

[5] Earnings call transcripts are sold and distributed by a handful of large data vendors, such as FactSet, S&P Global, and Refinitiv. They are also publicly available at https://seekingalpha.com and www.fool.com.

[6] www.bloomberg.com, www.wsj.com, www.reuters.com, www.barrons.com, www.nytimes.com, www.cnbc.com, www.marketwatch.com, www.ft.com, https://finance.yahoo.com, https://apnews.com, www.cnn.com, www.foxnews.com, www.foxbusiness.com.

Exhibit 2. Information Flow by Company Size and Time of Year



Note: Data are for 1 January 2019 to 31 December 2021.

Exhibit 3. Example Usage of Diffbot Query Language

Plain English Request	I want to see every news article about Nike or Adidas, written in English, published by Reuters, Yahoo! Finance, or Bloomberg, and published between the years of 2016 and 2021.

Diffbot Query Language (DQL)	`type:Article`
	`OR(tags.uri:"EnpdllhyQM-SckYV3z0i1Dw",tags.uri:"EikhmjkE8MlSihDC4RKNRWg")`
	`language:"en"`
	`OR(pageUrl:"reuters.com",pageUrl:"finance.yahoo.com",pageUrl:"bloomberg.com")`
	`date>="2016-01-31" date<="2021-12-31"`

interfaces (APIs) remove the barriers of collecting, cleaning, and structuring data, so the analyst can remain focused on investment insights.

A Robo-Analyst's First Steps: Word Embeddings and Text-Based Analytics

Having defined the "who" with entity mapping, the next step is identifying "what" topics are being discussed in a document, paragraph, or sentence. Counting the frequency of specific words in a body of text is a simple way to start, but doing so necessitates determining all the possible words and combinations of words related to each topic. Determining synonyms by committee can be very time consuming and requires deep subject matter expertise. For example, research on terms related to "greenhouse gas" would include "GHG," "chlorofluorocarbon," "carbon emissions," individual greenhouse gases, such as "methane" and "carbon dioxide," and so on.

To solve this problem of synonyms and tangentially related terms, word embeddings, a foundational NLP concept, can find relevant keywords associated with a given topic. Once trained on a corpus of text, word-embedding models can generate a family of semantically similar phrases for any given seed word (the technical name for the word the researcher inputs to find related terms). With these topic definitions, commonly referred to as "dictionaries," the researcher can begin to comb volumes of text in a comprehensive, consistent, and scalable manner.

Process overview: Identifying ESG topics in text

The process for connecting specific topics to companies follows the sequence below (see **Exhibit 4**), which was informed by research from the academic community and investment managers:[7]

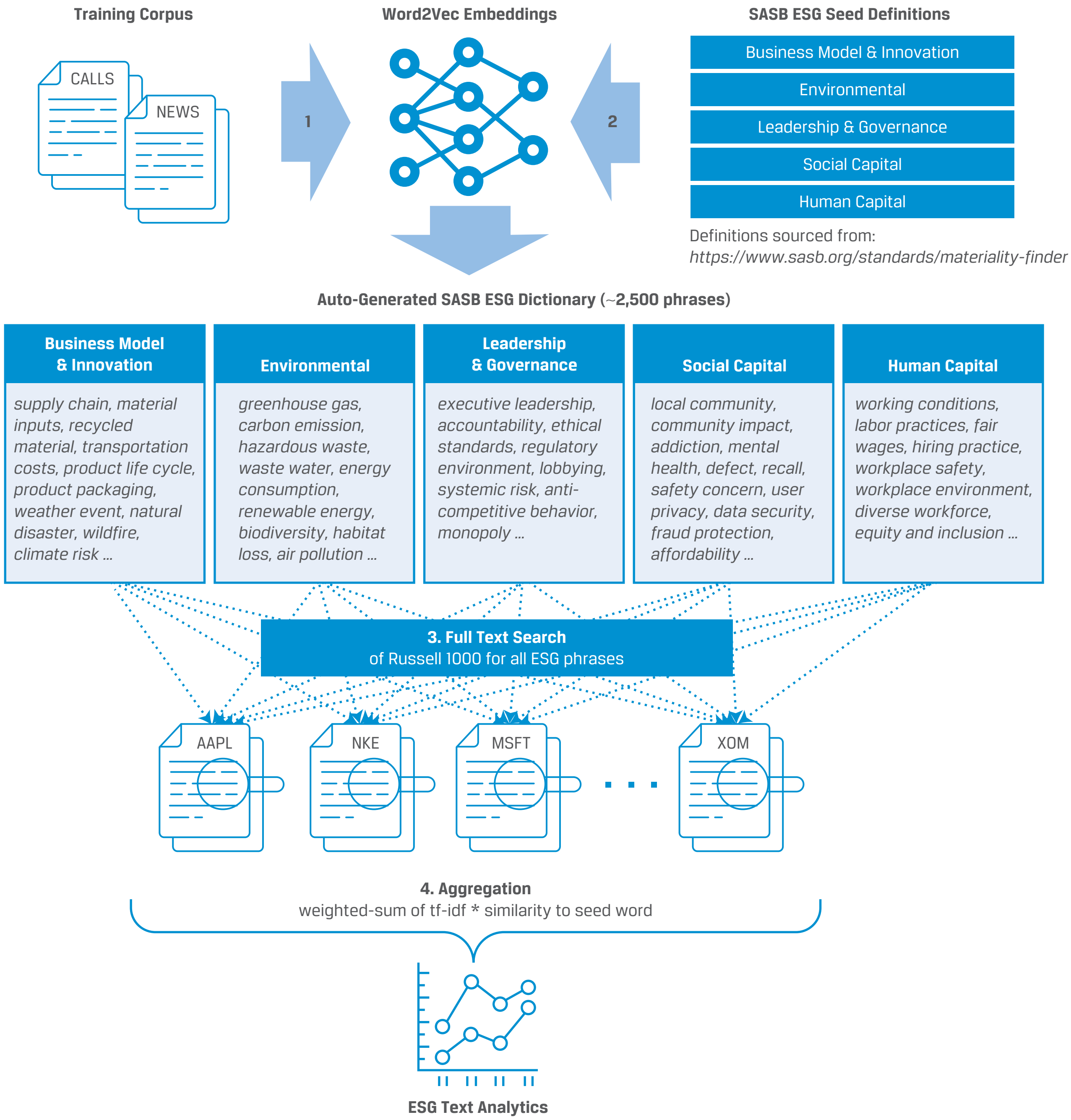
1. Train a word2vec (word-embedding) model on earnings call transcripts and news articles.
2. Feed the model a small set of seed words for which the researcher is seeking synonyms and related words. The Sustainability Accounting Standards Board (SASB), a pioneer in determining ESG standards, has defined 26 ESG categories.[8] This study uses roughly two- to three-word summaries for each of the 26 categories as seed phrases for the word-embedding model, which then created 26 phrase lists (dictionaries), resulting in approximately 2,500 phrases across all dictionaries.
3. Match words in the ESG category–specific dictionaries with instances of those phrases in the corpus (news articles and earnings calls).
4. Finally, the number of matches with terms in the dictionary needs to be rolled up to numerically define the original seed word. To do so, the researcher calculates the weighted sum of the phrase frequencies for each SASB issue. The weights are a combination of the term frequency–inverse document frequency (TF–IDF)[9] of a given phrase multiplied by its similarity to the seed phrase.

[7] This process is motivated by the creative work of Li, Mai, Shen, and Yan (2020) and Wu (2021).

[8] Seed words were manually extracted from issue-level descriptions available at the SASB's "Materiality Finder" website: www.sasb.org/standards/materiality-finder/?lang=en-us.

[9] TF–IDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The TF–IDF value increases proportionally with the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps adjust for the fact that some words appear more frequently in general.

Exhibit 4. Process Diagram: Applying Word-Embeddings to ESG


Training Corpus
CALLS
NEWS
1
Word2Vec Embeddings
2
SASB ESG Seed Definitions
Business Model & Innovation
Environmental
Leadership & Governance
Social Capital
Human Capital
Definitions sourced from:
https://www.sasb.org/standards/materiality-finder
Auto-Generated SASB ESG Dictionary (~2,500 phrases)
Business Model & Innovation
supply chain, material inputs, recycled material, transportation costs, product life cycle, product packaging, weather event, natural disaster, wildfire, climate risk …
Environmental
greenhouse gas, carbon emission, hazardous waste, waste water, energy consumption, renewable energy, biodiversity, habitat loss, air pollution …
Leadership & Governance
executive leadership, accountability, ethical standards, regulatory environment, lobbying, systemic risk, anti-competitive behavior, monopoly …
Social Capital
local community, community impact, addiction, mental health, defect, recall, safety concern, user privacy, data security, fraud protection, affordability …
Human Capital
working conditions, labor practices, fair wages, hiring practice, workplace safety, workplace environment, diverse workforce, equity and inclusion …
3. Full Text Search
of Russell 1000 for all ESG phrases
AAPL
NKE
MSFT
. . .
XOM
4. Aggregation
weighted-sum of tf-idf * similarity to seed word
ESG Text Analytics

This dataset now has a variety of use cases, including evaluating market trends and shifts in materiality for a given sector and monitoring company-specific news across a large swath of ESG issues. Also, the process itself is fully customizable to any set of seed words, allowing industry practitioners to update or adjust the original SASB taxonomy or create an entirely new taxonomy based on customized organizing principles.

ESG word embeddings

In 1957, a leading English linguist named John Rupert Firth (1957) noted that "you shall know a word by the company it keeps." Firth's prescient words have stood the test of time; neural networks can derive meaning by piecing together relationships between various combinations of different words in multiple contexts.

In word-embedding models, each word is represented by a numeric vector based on the usage of words and their co-occurrences over a large set of documents. As such, words used in similar ways are captured as having similar representations and meanings. For example, if the words that surround "carbon" are often related to the words that surround "emissions," the geometric distance between these word vectors will be small, inferring that "carbon" is closely related to "emissions." At the other extreme, if the context surrounding "carbon" is quite different from the context surrounding "hiring," the vector distance is large and the concepts are considered less related.

Training word2vec on earnings calls and news

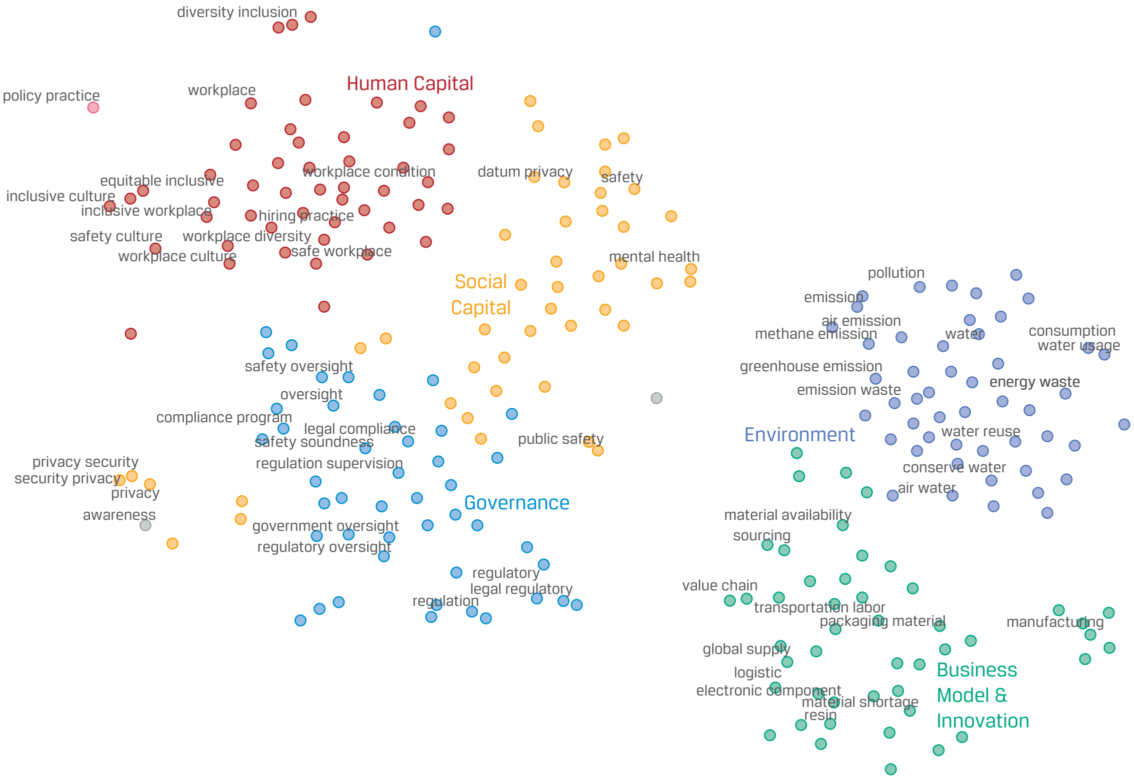
This study applied one of the most popular word-embedding algorithms, word2vec, invented by Tomas Mikolov and his team at Google in 2013. A decade in technology years might as well be a century in calendar years, yet word2vec's simplicity and effectiveness are still useful to researchers 10 years later.

The ESG word2vec model was trained on the earnings call and news data. Training a word2vec model can be done via a variety of open-source programming languages,[10] including Python, R, C, C#, and Java/Scala. The implementation presented here was carried out by using Python (the most popular language for implementing machine learning and NLP algorithms) within the topic-modeling library Gensim.[11]

The ecosystem of ESG topics, issues, and words

A common technique for visualizing and exploring similarities of words in a word2vec model is *t*-SNE,[12] which collapses multidimensional word vectors down to two dimensions. This approach reasonably maintains the relationships between words, while making the dataset neatly viewable in a two-dimensional scatterplot. **Exhibit 5** illustrates where each of the most prominent SASB words appear relative to each other.

Exhibit 5. The ESG Ecosystem: t-SNE Visualization of ESG Word-Embeddings



[10] See Wikipedia's article on word2vec: https://en.wikipedia.org/wiki/Word2vec.

[11] Gensim is an open-source library for unsupervised topic modeling, document indexing, retrieval by similarity, and other NLP functionalities.

[12] *t*-Distributed stochastic neighbor embedding (*t*-SNE) is a method for visualizing high-dimensional data by giving each datapoint a location in a two- or three-dimensional map and doing so in such a way that similar objects are modeled by nearby points and dissimilar ones by distant points.

The output in Exhibit 5 demonstrates intuitive relationships between SASB's five major ESG topics. For example, human capital and social capital are displayed near one another, meaning their underlying word embeddings often share similar contexts and themes within the financial corpus. Similarly, governance has overlapping language with social and human capital. Looking at the subtopics in black font, some governance issues, such as "safety," tend to relate to employees, customers, and public perceptions, but terms tied to industry standards, such as "regulation" and "oversight," are more distant from human and social capital.

Even a less experienced ESG analyst could replicate the overlap and interrelationships of ESG terminology in an approximately similar way. However, in this example, a machine with no prior information learned these relationships in about two hours of computation time. This trained word-embedding model is repeatable and consistent and provides a foundation for unleashing endlessly scalable "machine knowledge" onto a wide variety of contexts and information sources.

Text-based investment insights

The word embeddings illustrated in Exhibit 5 provide an intuitive map of individual words to ESG. These terms, which are collected into dictionaries for each ESG theme, are then detected and tabulated across millions of news articles and earnings calls. ESG themes (the "what") can now be tied to individual companies (the "who"). The output of this process can be considered a metric for the ESG concept of materiality, which measures the significance of a given ESG topic to a sector or company.

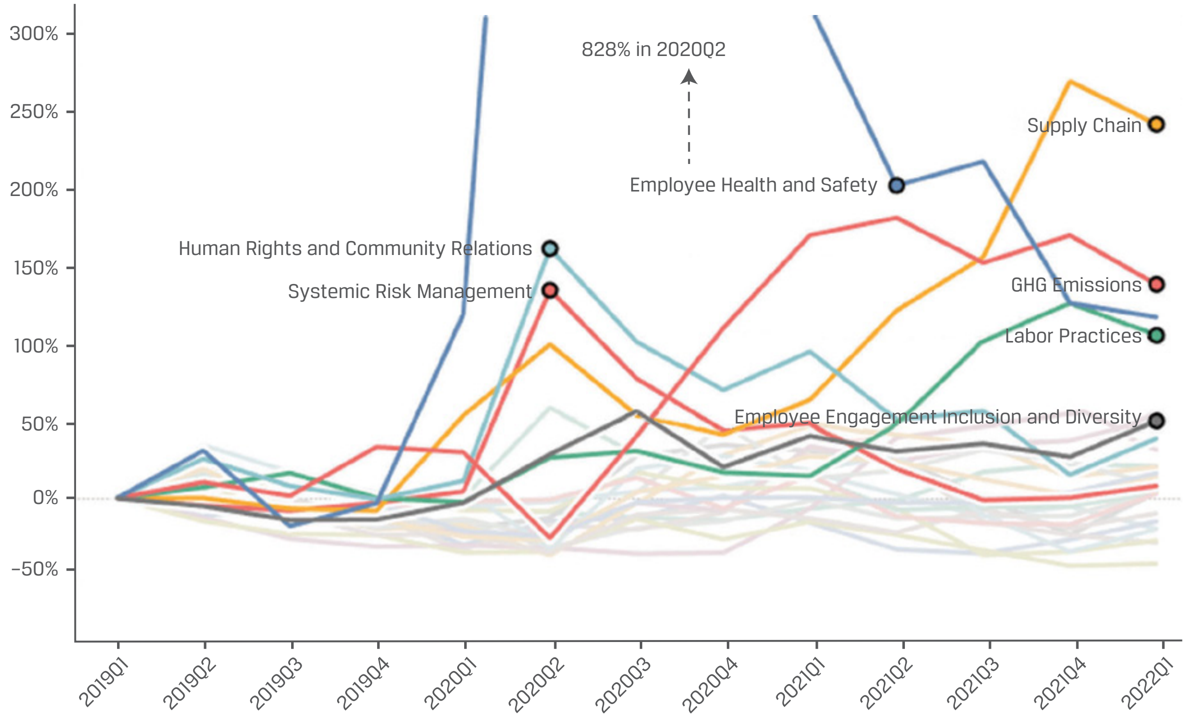
Guided by large teams of sustainability experts, materiality mapping is primarily qualitative and relatively static in standardized ESG frameworks. By contrast, the text-based model in this study systematically and dynamically connects companies to their most relevant ESG issues over time. This is particularly relevant for keeping up with shifts in corporate strategy. For example, Amazon's ESG materiality evolved as the company transitioned from selling books online to cloud computing, food distribution, and health care. Mergers and acquisitions can also abruptly shift materiality.

These dynamic, company-level ESG exposures can then be rolled up and applied to a variety of investment use cases, including evaluating market trends, identifying shifts in materiality for a given sector, creating thematic baskets of stocks focused on certain ESG issues, and monitoring company-specific news across a large swath of ESG risks and opportunities.

Trend analysis

Applying this framework at the highest level, **Exhibit 6** illustrates what mattered most in ESG for the Russell 1000 from 2019 to 2022, highlighting ESG issues that created the most chatter in the financial press and company earnings calls.

Exhibit 6. Trend Analysis of ESG Issues



Interpreting the output in Exhibit 6, employee health concerns immediately spiked at the onset of the pandemic, well ahead of the imposition of public health restrictions in early 2020 (AJMC 2021). Shortly thereafter, systemic risk management rose in frequency. Thereafter, the NLP-based materiality measure detected widespread supply chain issues due to abrupt shifts in supply and demand. Exhibit 6 also shows a 50% increase in references to labor practices in 2021. This increase was later confirmed by the National Labor Relations Board (NLRB), which reported a 58% increase in the number of union representation petitions over that same year (NLRB 2022). With no a priori knowledge beyond basic seed words, this aggregate market perspective of ESG topics detected several significant market drivers in real time, unlike traditional ESG reporting outlets.

Dynamic materiality

Company-level ESG mappings can be rolled up for sector analysis in much the same way that company valuations, such as the price-to-earnings ratio (P/E), can be rolled up into sector valuations. In other words, this machine-driven process yields a materiality map (**Exhibit 7**) that is dynamic, consistent, and customizable, which connects sectors to their most prominent ESG issues over time.

The NLP process generated an intuitive and compelling materiality map. Among other insights, environmental topics are mapped to energy-intensive sectors, such as energy, materials, and utilities, while consumer-facing industries (staples and discretionary) are mapped to human and social capital.

Company-level alerts and profiles

The same NLP framework allows analysts to build ESG profiles for individual companies to identify, highlight, and monitor the most relevant company-specific ESG trends.

In **Exhibit 8**, the bars represent the aggregate frequency of a given ESG issue across thousands of news articles, broken out by year. This dashboard is intended to highlight company trends. If overlayed with a user interface, analysts could click on each individual bar to investigate specific statements from the underlying news articles and earnings calls.

As Exhibit 6 showed, labor issues are rising from an aggregate market-level perspective and tied closely to consumer-facing sectors via materiality mapping (Exhibit 7). Connecting those big-picture themes to the

Exhibit 7. Text-Based ESG Materiality Map

	Most Material	2nd Most	3rd Most	4th Most	5th Most
Energy	GHG Emissions (ENV)	Air Quality (ENV)	Water & Wastewater Mgmt. (ENV)	Ecological Impacts (ENV)	Critical Incident Risk Mgmt. (GOV)
Materials	Material Sourcing & Efficiency (BM&I)	Product Design and Life-Cycle Mgmt. (BM&I)	Water & Wastewater Mgmt. (ENV)	Air Quality (ENV)	Supply Chain (BM&I)
Industrials	Labor Practices (HC)	Waste & Hazardous Materials Mgmt. (ENV)	Material Sourcing & Efficiency (BM&I)	Air Quality (ENV)	Supply Chain (BM&I)
Utilities	Energy Management (ENV)	GHG Emissions (ENV)	Physical Impacts of Climate Change (BM&I)	Mgmt. of Legal and Regulatory Env. (BM&I)	Air Quality (ENV)
Real Estate	Ecological Impacts (ENV)	Business Ethics (GOV)	Access & Affordability (SOC)	Labor Practices (HC)	Systemic Risk Mgmt. (GOV)
Financials	Systemic Risk Mgmt. (GOV)	Critical Incident Risk Mgmt. (GOV)	Mgmt. of Legal and Regulatory Env. (GOV)	Physical Impacts of Climate Change (BM&I)	Business Ethics (GOV)
Consumer Staples	Labor Practices (HC)	Supply Chain (BM&I)	Material Sourcing & Efficiency (BM&I)	Ecological Impacts (ENV)	Product Design & Life-Cycle Mgmt. (BM&I)

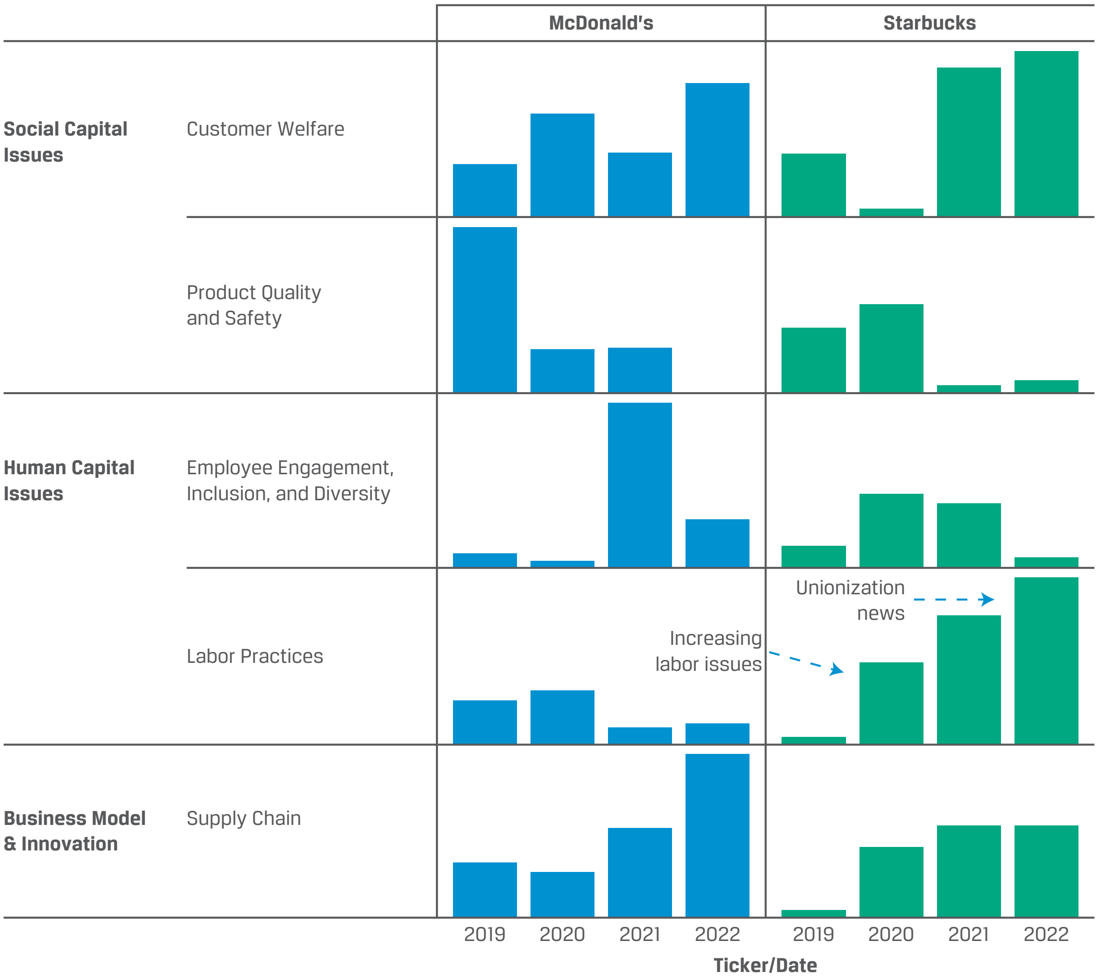
(*continued*)

Exhibit 7. Text-Based ESG Materiality Map (*continued*)

	Most Material	2nd Most	3rd Most	4th Most	5th Most
Consumer Discretionary	Access & Affordability (SOC)	Labor Practices (HC)	Supply Chain (BM&I)	Employee Health & Safety (HC)	Selling Practices & Product Labeling (SOC)
Information Technology	Data Security (SOC)	Customer Privacy (SOC)	Supply Chain (BM&I)	Selling Practices & Product Labeling (SOC)	Product Design & Life-Cycle Mgmt. (BM&I)
Communication	Customer Privacy (SOC)	Competitive Behavior (BM&I)	Human Rights & Community Relations (SOC)	Employee Engagement Inclusion & Diversity (HC)	Selling Practices & Product Labeling (SOC)
Health Care	Product Quality & Safety (SOC)	Customer Welfare (SOC)	Selling Practices & Product Labeling (SOC)	Access & Affordability (SOC)	Competitive Behavior (GOV)

Note: The SASB legend is environmental = (ENV), leadership & governance = (GOV), business model & innovation = (BM&I), human capital = (HC), and social capital = (SOC).

Exhibit 8. Company-Level ESG Profiles: McDonalds vs. Starbucks



company level, two of the largest public-facing brands in the world, Starbucks and McDonald's, can be scrutinized along a handful of their most material ESG issues.[13]

Starbucks' labor practices received additional attention at the onset of COVID-19 when many companies faced staff shortages and wage pressure. Exhibit 8 indicates that labor issues were already brewing in 2020, culminating in the announcement of a large minimum wage increase in 2021 and a coordinated push for unionization in May 2022. By contrast, McDonald's faced less scrutiny on labor practices; however, supply chain durability issues, another hidden risk that arose from the economic disruption of COVID-19 (Exhibit 6), appeared to generate more attention given the company's supply chain challenges in 2022.

Training Techniques: Language Models and NLP Tasks

The process described in the subsection titled "Process overview: Identifying ESG topics in text" dynamically and systematically captures fast-moving, real-time shifts in ESG events at the market, sector, and company levels. For best results, however, the underlying word-embedding model needs to be monitored and periodically retrained to capture broader shifts in the English language. This section provides real-world examples highlighting this point, while also clarifying key differences between domain-specific training (e.g., word embeddings derived from financial text) and task-specific training (topic classification of ESG issues). This distinction helps frame more advanced NLP models discussed in subsequent sections of this chapter.

Retraining Models for the Changing Linguistic and Cultural Landscape

As the world changes, so too do the words used to describe it. Word-embedding models learn from the language they are trained on. Therefore, as the language shifts, models trained on older versions of text will lose efficacy when applied to current issues.

Language shifts can be illustrated by analyzing the topic of "telecommuting" across similar word-embedding models that were trained at different times. Telecommuting, an SASB Human Capital topic focusing on workplace standards and employee well-being, underwent a monumental cultural shift during the pandemic. **Exhibit 9** compares how three word-embedding models—Google (created in 2013), Facebook (created in 2017), and the custom model built for this paper (in 2022)—determine which phrases are most closely related to "telecommuting."

Looking closer at the actual terms related to telecommuting, the older models from Google and Meta reflect a different lexicon with more emphasis on ways to maximize office time and minimize commuting time, whereas this study's model, which is trained on language from the last few years, relates telecommuting to only remote-work concepts and picks up the pandemic-driven paradigm shift toward working from home. While pretrained, open-source models are effective for exploring new ideas and deploying minimally viable products, outdated language models, no matter how accurate at the time of training, will fail to identify significant shifts in meaning.

Exhibit 9. Top 5 Phrases Most Similar to "Telecommuting"

Source	Google	Meta	This Study
Year	2013	2017	2022
Model	word2vec-google-news-300	fasttext-wiki-news-subwords-300	word2vec-newscalls-cfa
1	Teleworking	Teleworking	Remote work
2	Flextime	Commuting	Work from home
3	Compressed workweeks	Flextime	Teleworking
4	Flextime schedules	Job sharing	Shift to remote
5	Carpooling	Work at home	Work remotely

[13] Starbucks' material issues can be found at the SASB's Materiality Finder webpage: www.sasb.org/standards/materiality-finder/find/?company[]=US8552441094&lang=en-us. Those for McDonald's can be found at www.sasb.org/standards/materiality-finder/find/?company[]=US5801351017&lang=en-us.

Domain-Specific Training versus Task-Specific Training

There are two key steps for training a model to accomplish a specific NLP task. First, it needs to learn the language of a specific domain (financial text). Second, it needs to apply that learned knowledge to accomplish a specific task (identify what a sentence is about or how a topic is described).

In the subsection titled "Process overview: Identifying ESG topics in text," the word-embedding model (word2vec) trained itself on text from financial news and earnings calls, allowing it to better understand language in the financial domain. This training step was an example of unsupervised learning, meaning the algorithm learned patterns and associations directly from the text by itself without any need for human input.

Even if a model understands the language of a given domain, it still needs the requisite knowledge to solve specific tasks. The next step in the process—creating dictionaries from seed words—was an example of semi-supervised learning because the human researcher defined seed words to describe SASB's ESG categories that the machine then used to create phrase dictionaries of synonyms and related words, which were used to find topic matches in financial articles and earnings calls. While lacking sophisticated machine knowledge, this semi-supervised approach is perfect for effectively and efficiently exploring any theme (ESG or otherwise).

For more narrowly defined classification tasks, researchers often apply supervised learning, which requires humans to manually label many thousands of sentences with predefined classes, or categories. The machine then learns from these labels, detecting underlying patterns and relationships, and applies that knowledge to classify entirely new sentences that it has never seen before. While supervised learning is more time consuming and labor intensive, classification outputs tend to be more targeted and precise than unsupervised models with no human input.

The next two sections—a case study from Robeco and a section on sentiment classification with large language models—illustrate how this two-step process can be applied to real-world ESG problems: first by establishing requisite domain-specific knowledge via unsupervised learning and then by applying supervised learning to solve specific classification tasks learned from the intelligence of human labels.

NLP at Robeco: An Industry Case Study

Mike Chen and the Alternative Alpha Research Team at Robeco authored the research discussed in this section, which demonstrates the power of word embeddings and machine learning (ML) for NLP tasks, weaving the United Nations Sustainable Development Goals (SDGs) into ESG analysis.

NLP can be used for sustainability and alpha purposes, such as detecting employee sentiment from comments in online forums or determining traits of corporate culture emphasized by senior management. In this example, Chen, Mussalli, Amel-Zadeh, and Weinberg (2022) demonstrate how NLP can be used to measure corporate alignment with the 17 SDGs.

In recent years, the SDGs have emerged as a framework that impact-minded investors can use to align their investment portfolios with sustainability. The SDGs are defined qualitatively, which raises questions as to how company products and operations can be quantitatively measured in alignment with each of the goals. NLP is one way to translate qualitative descriptions into quantitative scores. Chen et al. (2022) described a combination NLP and ML framework to do so, illustrated in **Exhibit 10**.

In this framework, the authors outline a two-step approach whereby the initial inputs into the NLP model are various company CSR reports. These are vectorized in Step 1 via word embedding, which allows the model to understand individual words and the context in which the words appear, thereby obtaining deeper insight into the written text.

In Step 2, the output vectors are passed into the ML portion of the framework, which has been fed labels showing company alignment with the specific SDGs. The vectors, in conjunction with the SDG alignment labels, allow ML to ascertain how vectorized descriptions of business models and products map to the various SDGs (an example of supervised learning). This two-step framework provides significant flexibility, because each component (the corpus, the labeling inputs, and the NLP and ML algorithms) can be substituted as newer technologies are developed and better or more appropriate corpus or alignment data become available.

A key challenge when classifying SDG alignment using ML is that few companies are aligned with any given SDG. This data imbalance can present a problem for ML algorithms because the naive approach of classifying all samples to the majority case (in our example, the majority case is non-alignment to any SDG) can result in reasonably high accuracy. To address this issue, the authors used the synthetic minority oversampling technique (SMOTE) to artificially create training samples to allow the ML algorithm to learn on a balanced training set.

Since the data sample is imbalanced, the authors also used recall, precision, and F1 score to measure and compare the performance of various permutations of the NLP–ML framework. The authors found the best results are achieved using a combination of the doc2vec embedding technique for the NLP portion and support vector machine (SVM) for the ML part of the framework. The results obtained are shown in **Exhibit 11**. Under this combination, the F1 scores for alignment to the various SDGs range from a low of 69% to a high of 83%.

This case study demonstrates how NLP and ML can be combined to improve the investment decision-making process while maintaining alignment with clients' unique sustainability interests.

Exhibit 10. NLP–ML Framework Used to Assess a Company's Alignment with UN SDGs

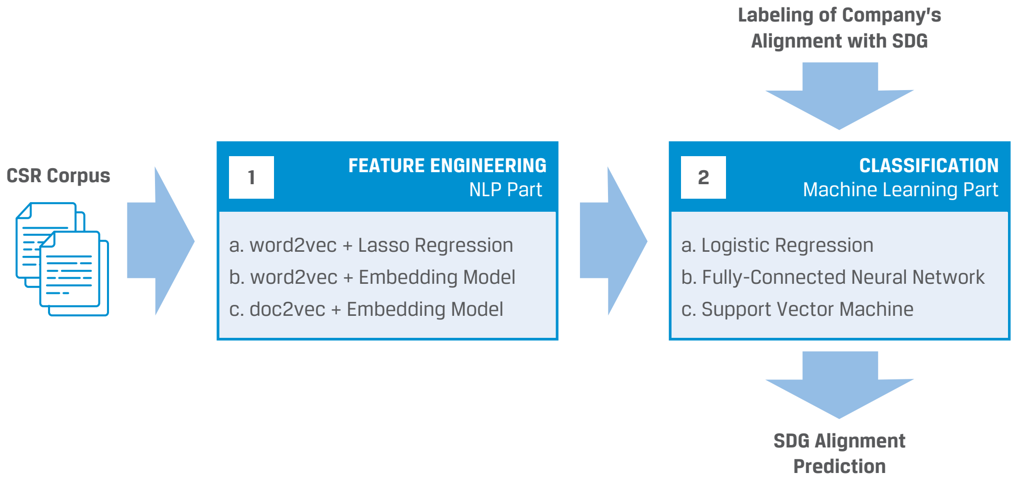

Labeling of Company's Alignment with SDG
CSR Corpus
1
FEATURE ENGINEERING
NLP Part
a. word2vec + Lasso Regression
b. word2vec + Embedding Model
c. doc2vec + Embedding Model
2
CLASSIFICATION
Machine Learning Part
a. Logistic Regression
b. Fully-Connected Neural Network
c. Support Vector Machine
SDG Alignment Prediction

Exhibit 11. Classifier Accuracy Measures

	Balanced Test Acc.	F1 Score	Recall	Precision
SDG1	0.87	0.70	0.81	0.62
SDG2	0.83	0.70	0.67	0.73
SDG3	0.79	0.80	0.74	0.86
SDG4	0.65	0.73	0.61	0.90
SDG5	0.67	0.76	0.66	0.89
SDG6	0.79	0.78	0.83	0.74
SDG7	0.78	0.79	0.76	0.82
SDG8	0.76	0.76	0.66	0.89
SDG9	0.81	0.81	0.83	0.79
SDG10	0.64	0.72	0.67	0.78
SDG11	0.88	0.73	0.82	0.66
SDG12	0.77	0.83	0.76	0.90
SDG13	0.79	0.77	0.67	0.92
SDG14	0.86	0.81	0.78	0.85
SDG15	0.76	0.79	0.71	0.89
SDG16	0.87	0.69	0.84	0.59

Context Is King: Large Language Models and Sentiment Analysis

Having covered "who" is being discussed (entity mapping) and "what" (insights from matching words in the corpus with automatically generated ESG dictionaries), the next step addresses "how" a given topic is discussed in the text. The company-specific dashboard in Exhibit 8 highlighted how often company-specific ESG issues were mentioned in the news for McDonald's and Starbucks. While helpful, the tone of the articles was absent from that analysis. Were events and topics recounted in a positive or negative light? Is the company doing better or worse on a specific ESG topic? Sentiment analysis offers answers to these questions by measuring the magnitude of an ESG event, as well as the tone and shifts in tone over time.

Beyond Word Embeddings

While word2vec models serve a wide variety of applications, they have a critical limitation: context awareness. In word2vec, each word is represented by a single vector, resulting in a single meaning. For example, the following two sentences use the word "freeze" in different contexts that change the underlying meaning:

- With demand slowing, the firm is undergoing a hiring *freeze*.
- An unseasonable mid-spring *freeze* impaired the coffee harvest.

Using a word2vec-based approach, the word "freeze" will maintain the same vector representation despite the obvious differences in context. The model's inability to capture these differences in meaning will have a negative impact on the accuracy of downstream classification tasks.

BERT and the leap from words to sentences

In 2018, the AI language team at Google created Bidirectional Encoder Representations from Transformers (BERT), which was a clear inflection point in NLP. The large language model (LLM), which learned much of the nuanced meaning of the English language from Wikipedia (approximately 2.5 billion words) and Google's Book Corpus (approximately 800 million words), vastly improved the accuracy of NLP models by enhancing contextualization.[14]

LLMs, such as BERT, allow for multiple representations of each word, depending on the other words in a sentence and the order in which they appear. Operating on sentences as inputs instead of individual words or phrases, these new models brought machines one step closer to understanding text in a nuanced, human-like way (see, for example, **Exhibit 12**).

Portable intelligence with transfer learning

A massive dataset of 3.3 billion words contributed to BERT's vast knowledge of the English language, but this immense scale also limited its adoption beyond an exclusive set of AI superlabs with enormous computational resources. Retraining LLMs for every new task was impractical, expensive, and completely out of reach for everyone else.

Around the same time, practitioners began combining LLMs with transfer learning, a complex mathematical technique that enables the vast knowledge inherent in such models to be transferred to a new classification task. With this new ML architecture, a general purpose model, such as BERT, could be pretrained, packaged, and reused as a starting point for fine-tuning domain-specific tasks, such as sentiment classification of financial text. This advancement allowed research teams of all sizes to inherit and apply existing knowledge, ultimately leading to rapid innovation.

Exhibit 12. Limitations of Word-Matching in Sentiment Analysis

	Sentiment via Word Matching	Actual Sentiment
"As far as we can tell, the company has yet to make any *progress* in this direction."	Positive	Negative
"In total, a lower nickel price eases some of our *concerns* about ATI and its metal (stainless/nickel alloys) exposures."	Negative	Positive

[14] Before BERT, a number of models improved accuracy across a battery of language tasks, including sentiment analysis. A few notable models include GloVe (in 2013), fastText (in 2016), CoVe (in 2017), ULMFiT (in 2018), ELMo (in 2018), and GPT-1 (in 2018).

Additionally, the training of LLMs requires roughly five times the carbon emissions compared with owning a car in the United States for one person's entire lifetime (Hao 2019). Transfer learning architecture boasts an improved environmental footprint by reducing the energy usage necessary for applying LLMs[15] (see **Exhibit 13**).

Democratizing NLP: Hugging Face transformers, PyTorch, and TensorFlow

Open source is the final ingredient necessary for the broad application of NLP. Hugging Face is at the forefront of an increasingly democratized NLP movement with its intuitively designed API that abstracts powerful deep learning libraries, such as PyTorch and TensorFlow,[16] thereby streamlining and simplifying ML workflows. The Hugging Face Hub, also known as "the GitHub of machine learning," allows its community members (from hobbyists to AI superlabs) to share their work with the world, hosting over 50,000 models, 5,000 datasets, and 5,000 demos.

Combining Topic and Tone

The three advancements described in the prior subsections (large language models, transfer learning, and Hugging Face transformers) can now be directly leveraged to create more nuanced ESG models that understand both the topic being discussed (the "what") and the tone (the "how").

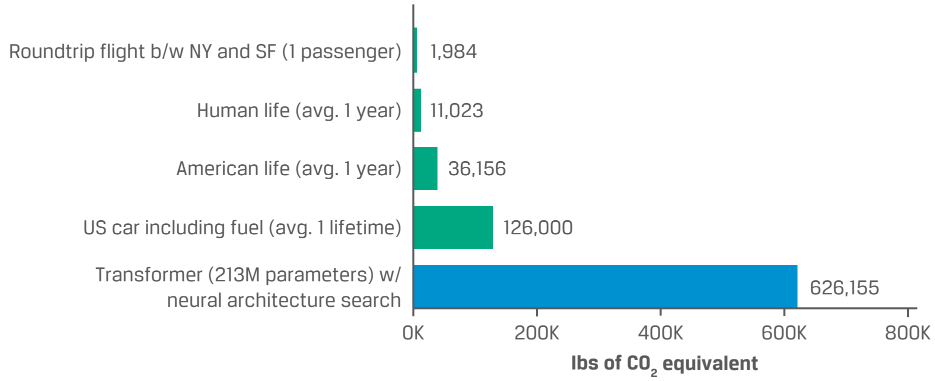
The next step of this study builds on work from researchers at Hong Kong University of Science and Technology (HKUST), who pre-trained the original BERT model on financial text using 10Ks and 10Qs (2.5 billion tokens), earnings calls (1.3 billion tokens), and analyst reports (1.1 billion tokens; Huang, Wang, and Yang, forthcoming). In other words, via Hugging Face's implementation of transformers, they were able to refine BERT (trained to the English language) to understand relevant financial language.

Huang et al. (forthcoming) also share two fine-tuned (task-specific) models: FinBERT-ESG and FinBERT-tone.[17] Both models were trained using supervised learning, with researchers manually labeling thousands of statements as "E," "S," or "G," (for FinBERT-ESG) and "positive," "negative," or "neutral" (for FinBERT-tone). This entire process, detailed in **Exhibit 14**, offers researchers an expedient way to combine topics with tone. Note, however, that while borrowing models is great for exploratory analysis, experimentation, and illustrative guides, for production, it is imperative to fully evaluate the labels and test overall efficacy for a specific NLP task.

Company-level alerting

To combine topic and tone, this study parses every sentence in the news dataset using the FinBERT-ESG classifier of Huang et al. for each sentence (E, S, G, or none) and then analyzes the same sentence using the FinBERT-tone classifier (positive, negative, or neutral). By aggregating the

Exhibit 13. Common Carbon Footprint Benchmarks (in pounds of CO_2 equivalent)



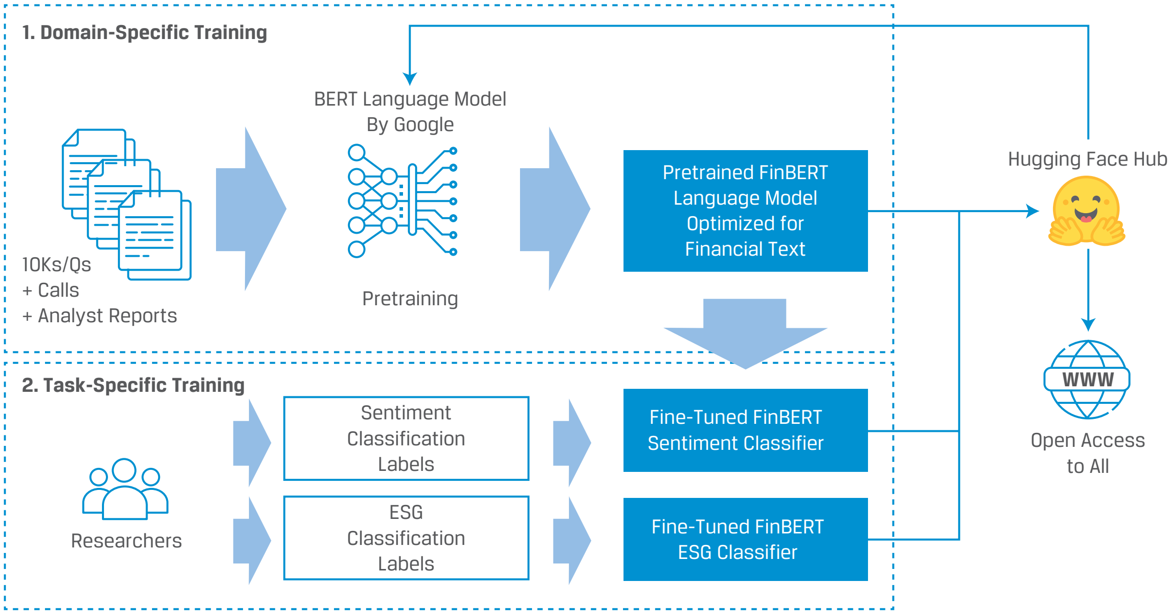
Sources: Hao (2019); data are from Strubell, Ganesh, and McCallum (2019).

[15] See Hugging Face's "How Do Transformers Work?" webpage at https://huggingface.co/course/chapter1/4.

[16] PyTorch and TensorFlow are free open-source ML and AI frameworks used most often for applications in NLP and computer vision. PyTorch was developed by Meta AI, and TensorFlow by Google.

[17] See the models' webpages at, respectively, https://huggingface.co/yiyanghkust/finbert-esg and https://huggingface.co/yiyanghkust/finbert-tone.

Exhibit 14. Process Diagram: Classifying ESG Topic and Tone with BERT


1. Domain-Specific Training
10Ks/Qs
+ Calls
+ Analyst Reports
BERT Language Model
By Google
Pretraining
Pretrained FinBERT Language Model Optimized for Financial Text
Hugging Face Hub
WWW
Open Access to All
2. Task-Specific Training
Researchers
Sentiment Classification Labels
ESG Classification Labels
Fine-Tuned FinBERT Sentiment Classifier
Fine-Tuned FinBERT ESG Classifier

outputs of these two classifiers for each company, shifts in tone from 2021 to 2022 for a given ESG topic are visible, thereby offering a real-time system for detecting rising ESG risk for all the companies in a given equity universe.

Exhibit 15 displays the sentiment shift relating to environmental practices for individual companies. Companies in the upper-left corner represent sharply deteriorating sentiment from 2021 to 2022, which indicates rising risk related to perceptions about their environmental stewardship. In this example, NextEra Energy (NEE), PepsiCo (PEP), Lockheed Martin Corporation (LMT), and Uber (UBER) all received increasingly negative attention across a swath of environmental issues.

Deeper analysis into this methodology reveals the challenges of desensitizing models to pervasive events, such as severe market crashes, pandemics, and wars, because economic shocks can add noise and false positives to the results. For example, Uber's negative news was related to the increased cost of gas and its effect on the bottom line for its ridesharing services, which was only tangentially related to environmental practices, whereas the other companies received more scrutiny that directly targeted their environmental practices.

Applications in investment management

The following list provides examples of how this technology is being applied in investment management.

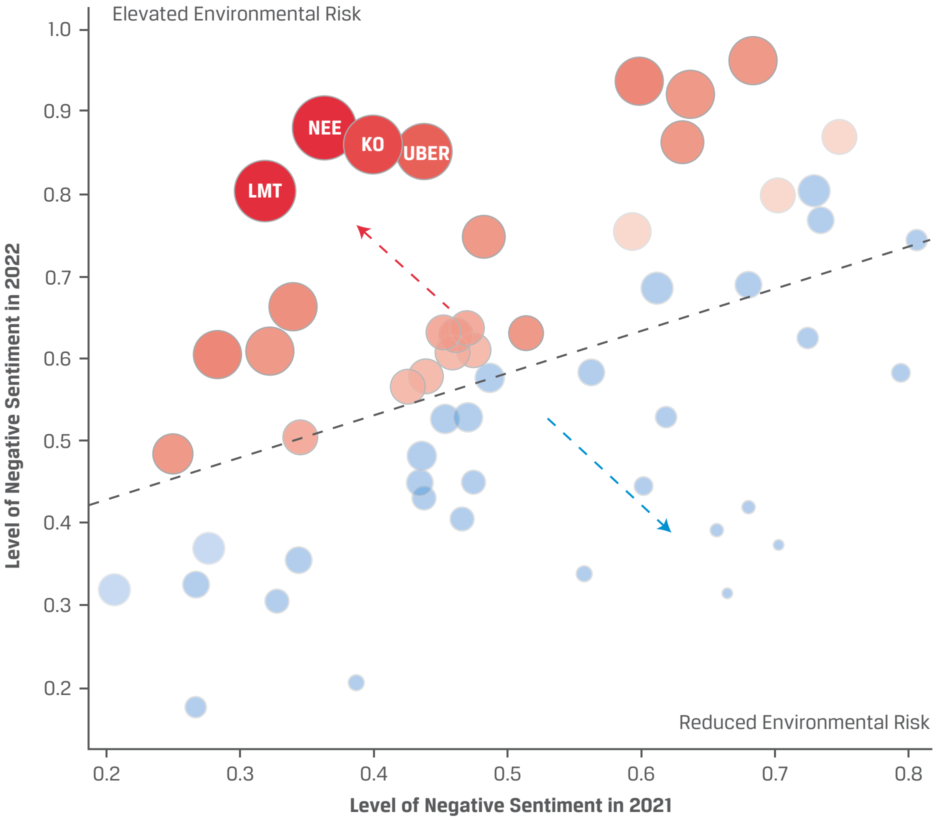
1. **Interactive dashboards:** Analysts tracking specific companies can create continually refreshed dashboards (as illustrated in Exhibit 15) and click on dots that link directly to relevant reports for further analysis.
2. **Risk alerts:** Risk alerting is a more automated and comprehensive approach to informing portfolio management decisions in real time by setting thresholds that send notifications to teams or combining real-time alerts with other measures, such as the application of an exclusionary stock screen.
3. **Quantitative analysis:** While the focus of this study is SASB-defined ESG issues, this framework is fully customizable to any company dimension (company culture scores, exposures to world events, etc.) covered in public data sources. Text-based signals are used across the industry as differentiating sources of alpha and for event-based risk control. This type of analysis is available in third-party applications, such as Alexandria Technologies, RavenPack (2022), RepRisk, and Truvalue, among many others.[18]

ESG Success: It Depends on Who You Ask

Comparing differences in tone between companies' earnings calls and their coverage in the financial press reveals notable if unsurprising results: Companies' descriptions

[18] For more information, go to www.alexandriatechnology.com/esg (Alexandria Technology); www.reprisk.com/ (RepRisk); and https://developer.truvaluelabs.com/catalog (Truvalue).

Exhibit 15. Text-Based ESG Alerting



of their ESG performance tend to be far more glowing than how the news characterizes their performance.

Exhibit 16 shows that management is far more likely to speak positively about the big three ESG categories than about non-ESG issues. The "exaggeration" or "spin" is most pronounced when management is discussing governance issues. Specifically, management's tone is nearly three times *more positive* when talking about governance (9.5 positive-to-negative ratio) than when discussing non-ESG issues (3.8 positive-to-negative ratio). News coverage offers the opposite perspective with similar magnitude: Governance issues are discussed three times *more negatively* compared to statements about non-ESG issues (0.5 versus 1.3, respectively).

Drawing Inference: Zero-Shot Classification and Greenwashing

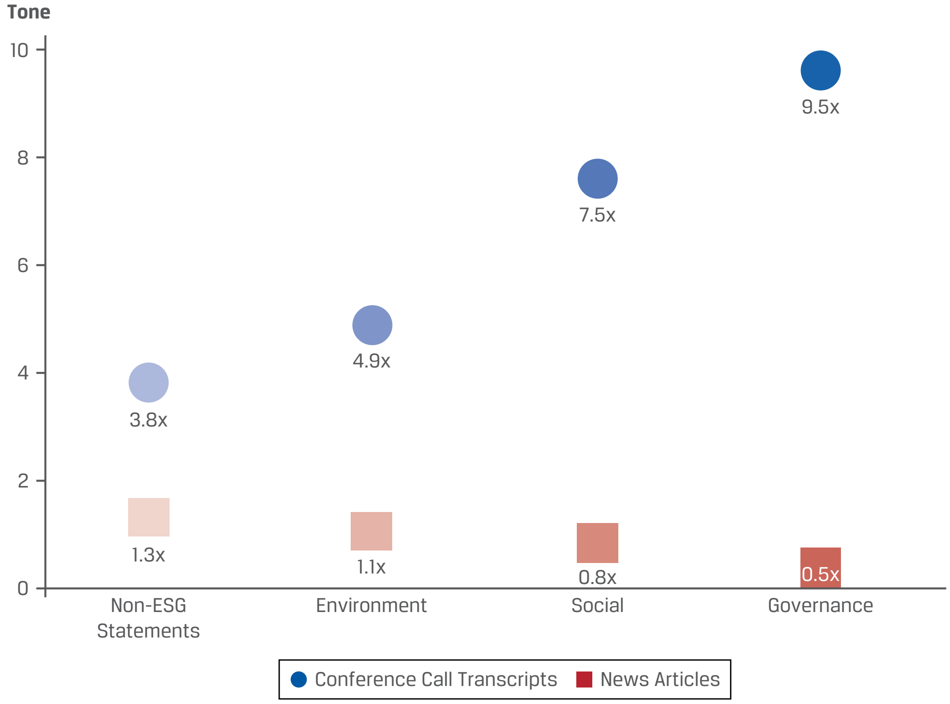
Sugar coating company performance in earnings calls is a well-known phenomenon to fundamental analysts and quants alike. A case study titled "Dissecting Earnings Conference Calls with AI and Big Data: American Century" in the report "AI Pioneers in Investment Management" (CFA Institute 2019, pp. 26–27) empirically demonstrates that NLP can identify signs of management spin, manipulation, and even obfuscation in earnings calls. In the context of ESG, there's a specific term for this phenomenon: greenwashing.

Greenwashing is an increasingly important topic in the investment community and society at large. It describes the activity of misleading the public about an entity's commitment and performance with respect to ESG topics. In fact, the negative implications of greenwashing might be the only common ground that exists between ESG investing's advocates and its skeptics. While most standardized ESG frameworks include more broadly defined issues, such as selling practices and product labeling, the narrow topic of greenwashing has yet to be explicitly measured, which presents an exciting opportunity to use NLP to supplement standardized ESG frameworks with metrics for greenwashing risk. Greenwashing increasingly carries heightened reputational and legal risk, so identifying early signals of this concept has substantial economic value.

Designing a Greenwashing Classifier

The goal of this section is to provide simple and effective techniques to identify sentences that contain content

Exhibit 16. Comparison of Tone between Conference Calls and News across Varying ESG Topics



Note: Tone = Ratio of positive-to-negative sentences.

related to greenwashing. A word search–based process—one that simply looks for the word "greenwashing" in text—is a reasonable place to start and would suffice for many use cases. But the model can be further refined with techniques that can classify sentences that are about greenwashing without explicit mention of the term in the text.

In this study, sentences flagged for greenwashing simply indicate that the content of a sentence is about greenwashing. This does not indicate that the company in question is greenwashing. Like the risk-alerting applications presented previously, classification techniques provide critical first steps for dynamically and systematically tracking high-stakes ESG issues in tandem with analyst oversight.

When a word has no synonyms

Eskimos have 50 words for "snow" because thriving over millennia in the frigid climate required a nuanced understanding of snow. By contrast, the ESG ecosystem that investors inhabit is so new that there is only one word that is typically used to describe greenwashing.

Applying the word2vec model to other business words with more history and nuanced meaning, such phrases as "data breach," yield a rich family of associated words, such as "hacked" and "security compromised." In contrast, the term "greenwashing" returns no dictionary of similar phrases because greenwashing is a relatively new term that combines a few concepts into one word with a meaning that can be roughly deconstructed as follows: greenwashing = company + misleading + sustainability.

To broaden the machine's understanding of greenwashing, one could apply a supervised learning approach wherein a group of researchers would label sentences ("about greenwashing" versus "not about greenwashing"), while attempting to address common label-quality issues related to ambiguity and differences in reviewer opinions. However, this would be an enormous task, especially considering that different sectors have entirely different materiality to ESG topics. The next section addresses the challenge of using unlabeled data and advances a foundational objective of AI research: maintaining or increasing model accuracy while reducing human input.

Introducing Zero-Shot Classification

Zero-shot classification is a relatively new NLP task that answers the question, *what is this sentence about?* Unlike fine-tuned classifiers, it does so without previously seeing a sentence or a label (hence the name "zero shot"). Hugging Face researchers found that zero-shot classifiers are surprisingly accurate when deployed on LLMs (Davison 2020), albeit not quite as accurate as classifiers explicitly trained on the same topic via thousands of human labels.

Powered by natural language inference

Zero-shot classifiers were popularized by Yin, Hay, and Roth (2019) and rely on natural language inference (NLI), an NLP task that compares two sentences—a premise and a hypothesis—and determines whether the hypothesis aligns with the premise (entailment), contradicts it (contradiction), or does neither (neutral). Our study applies the hypothesis "This is about greenwashing" to a series of sentences using Meta's *bart-large-mnli* model, the most popular zero-shot classifier on the Hugging Face Hub.[19] The model responds with its assessment of whether the hypothesis is true, false, or neutral for each sentence. The output ranges between 0% (not at all related to greenwashing) to 100% (entirely related to greenwashing).

Test driving the zero-shot classifier

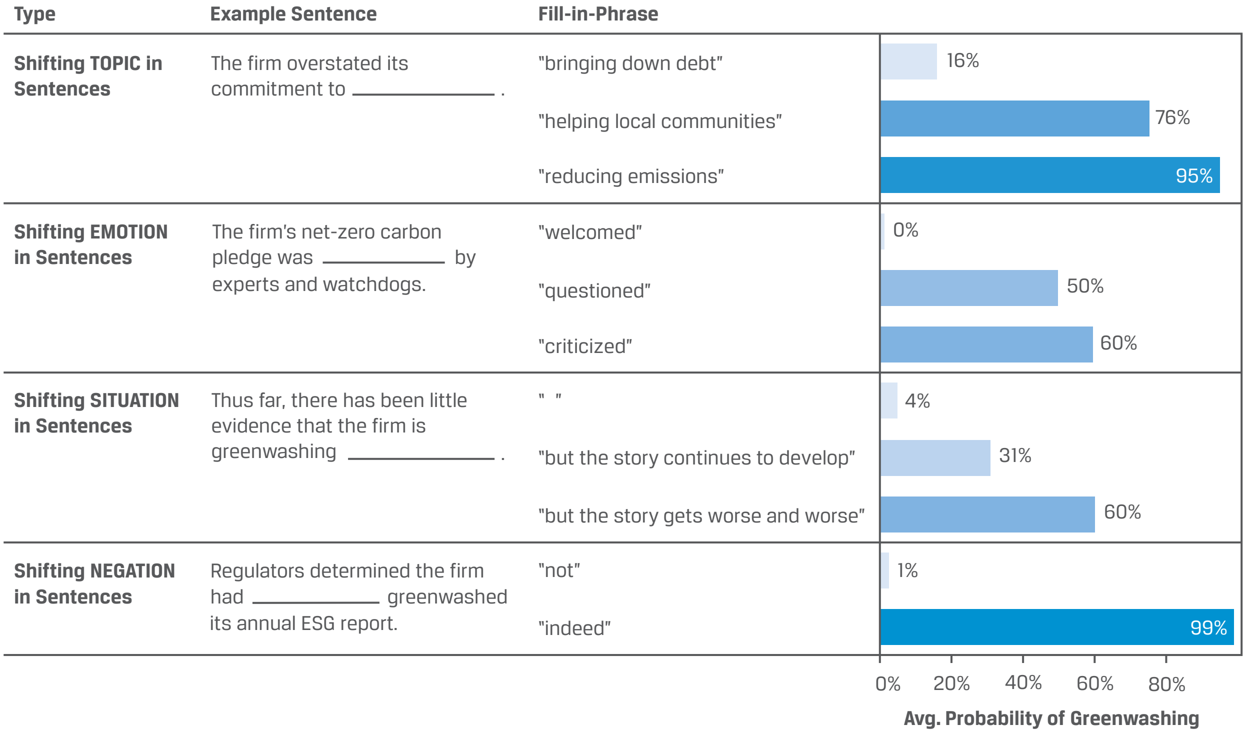
Zero-shot classifiers can understand multiple dimensions about a sentence—namely, topic, emotion, and situation (as described by Yin et al. 2019). This is especially appropriate for such a relatively novel and complex topic as "greenwashing." **Exhibit 17** illustrates the model's intuition and sensitivity to subtle changes in language. The zero-shot model intuitively adjusts its output (ranging from 0% to 100%) to the hypothesis "This sentence is about greenwashing" according to changes in the same sentence (relating to tone, emotion, situation, and negation).

Exhibit 17 demonstrates how the model handles nuanced meaning by substituting words in each sentence. For example, the sentence "The firm overstated its commitment to *debt*" (topic = debt) has a 16% chance of being about greenwashing, whereas the sentence "The firm overstated its commitment to *reducing emissions*" (topic = emissions) has a 95% chance of being related to greenwashing. This example-driven analysis highlights the model's ability to understand language and infer meaning in the context of ESG.

Preliminary analysis

Exhibit 17 helps illustrate how the model "thinks." The next step, detailed in **Exhibit 18**, examines *how well* it thinks by manually labeling 200 sentences and evaluating its ability to distinguish between three different types of sentences: sentences that contain the words "greenwash," "greenwashed," or "greenwashing" (15 sentences); sentences

Exhibit 17. Test-Driving the Zero-Shot Classifier



Type	Example Sentence	Fill-in-Phrase	Avg. Probability of Greenwashing
Shifting TOPIC in Sentences	The firm overstated its commitment to ______ .	"bringing down debt"	16%
		"helping local communities"	76%
		"reducing emissions"	95%
Shifting EMOTION in Sentences	The firm's net-zero carbon pledge was ______ by experts and watchdogs.	"welcomed"	0%
		"questioned"	50%
		"criticized"	60%
Shifting SITUATION in Sentences	Thus far, there has been little evidence that the firm is greenwashing ______ .	" "	4%
		"but the story continues to develop"	31%
		"but the story gets worse and worse"	60%
Shifting NEGATION in Sentences	Regulators determined the firm had ______ greenwashed its annual ESG report.	"not"	1%
		"indeed"	99%

[19] Go to https://huggingface.co/facebook/bart-large-mnli. The model was developed by Meta Platforms, downloaded roughly 1.5 million times per month in 2022, and trained on 433,000 hypothesis/premise pairs.

Exhibit 18. Preliminary Accuracy Measures for Zero-Shot Classification of "Greenwashing"

	Confusion Matrix 1 **Matching Word "Greenwash(ed/ing)"**		**Confusion Matrix 2** **Zero-Shot Classifier for "Greenwashing"**	
200 Total Sentences	Predicted: Sentences Related to Greenwashing	Predicted: Sentences Unrelated to Greenwashing	Predicted: Sentences Related to Greenwashing	Predicted: Sentences Unrelated to Greenwashing
Actual: Sentences Related to Greenwashing 50 Sentences	15 (True Positive)	35 (False Negative)	30 (True Positive)	20 (False Negative)
Actual: Sentences Unrelated to Greenwashing 150 sentences	0 (False Positive)	150 (True Negative)	0 (False Positive)	150 (True Negative)
	Precision = 100%; Recall = 30%; F1 Score = 46%		Precision = 100%; Recall = 60%; F1 Score = 75%	

that describe greenwashing without mentioning the term (35 sentences); and sentences completely unrelated to greenwashing (150 sentences). These 200 sentences were run through the zero-shot classifier using a 50% threshold; that is, if the classifier output is more than 50%, the sentence gets classified as relating to greenwashing. While this test is run on a small sample, it takes a critical first step in understanding the model's effectiveness.

The confusion matrixes in Exhibit 18 detail reasonably encouraging results. The rudimentary word-matching model is incapable of picking up on sentences that did not explicitly contain the terms "greenwash," "greenwashed," and "greenwashing." By contrast, the zero-shot classifier powered by *bart-large-mnli* correctly identified the 15 sentences containing those words, in addition to 15 additional sentences that merely alluded to greenwashing, all without generating any false positives. In other words, the zero-shot classifier captures twice the number of sentences that relate to greenwashing without making any additional mistakes.

Next steps

Running the same zero-shot classifier on a larger set of unlabeled data (approximately 5 million sentences from news) was fruitful but exposed the pain points of imbalanced classes (greenwashing is an extremely small fraction of broader market discussions), which requires further refinement of the original 50% threshold. There are also easily correctable blind spots in the classifier, such as the word "green" leading to false positives. These early insights provide a path forward for improving the model.

Tackling the subject of greenwashing head on with more scrutiny and analysis will ultimately benefit the public and improve on what constitutes adhering to ESG principles. Meanwhile, language models are advancing at an incredible pace, reducing the amount of human input required to achieve the same results. Improving accuracy while reducing costs is a powerful accelerator for additional research in this rich intersection of ESG and NLP.

Conclusion

The customizable and dynamic nature of NLP makes it an ideal match for the rapidly evolving data and definitions and the significant challenges of ESG analytics. The investment community's application of the techniques outlined in this chapter will further refine their usefulness, while the combination of humans and machines is fast becoming the new frontier in investment insight. There has never been a better time to collect, process, and draw meaning from text, and the innovations are only poised to continue.

Whether building AI tools from the ground up, sourcing solutions from third-party data providers, or effectively interpreting model outputs, the foundations detailed in this chapter can entirely reframe how problems are solved and redefine the questions that can be asked of data. Quite simply, NLP can move the investment industry beyond its current limitations of data, time, resources, and imagination and, in tandem with ESG, create meaningful and high-performing contributions to active investing.

We would like to thank Matt Spain, Rebecca Jackson, Vinod Chandrashekaran, Jonathan Berkow, Mike Chen, and Mike Tung.

References

AJMC. 2021. "A Timeline of COVID-19 Developments in 2020." (1 January). www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020.

Cerulli Associates. 2022. "ESG Issue." The Cerulli Edge—US Institutional Edition (First Quarter).

CFA Institute. 2019. "AI Pioneers in Investment Management." www.cfainstitute.org/-/media/documents/survey/AI-Pioneers-in-Investment-Management.pdf.

Chen, M., G. Mussalli, A. Amel-Zadeh, and M. Weinberg. 2022. "NLP for SDGs: Measuring Corporate Alignment with the Sustainable Development Goals." *Journal of Impact and ESG Investing* 2 (3): 61–81.

Davison, Joe. 2020. "Zero-Shot Learning in Modern NLP." *Joe Davison Blog* (29 May). https://joeddav.github.io/blog/2020/05/29/ZSL.html.

Firth, J. 1957. "A Synopsis of Linguistic Theory, 1930–55." In *Studies in Linguistic Analysis*, 1–31. Oxford, UK: Blackwell.

Hao, Karen. 2019. "Training a Single AI Model Can Emit as Much Carbon as Five Cars in Their Lifetimes." *MIT Technology Review* (6 June). www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/.

Henry, P., and D. Krishna. 2021. "Making the Investment Decision Process More Naturally Intelligent." Deloitte Insights (2 March). www2.deloitte.com/us/en/insights/industry/financial-services/natural-language-processing-investment-management.html.

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

IFC. 2004. "Who Cares Wins—Connecting Financial Markets to a Changing World" (June). www.ifc.org/wps/wcm/connect/topics_ext_content/ifc_external_corporate_site/sustainability-at-ifc/publications/publications_report_who-careswins__wci__1319579355342.

Li, Kai, Feng Mai, Rui Shen, and Xinyan Yan. 2020. "Measuring Corporate Culture Using Machine Learning." Working paper (29 June). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3256608.

NLRB. 2022. "Correction: First Three Quarters' Union Election Petitions up 58%, Exceeding All FY21 Petitions Filed." Office of Public Affairs (15 July). www.nlrb.gov/news-outreach/news-story/correction-first-three-quarters-union-election-petitions-up-58-exceeding.

RavenPack. 2022. "ESG Controversies: March 2022" (5 April). www.ravenpack.com/blog/esg-controversy-detection-march-2022.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. "Energy and Policy Considerations for Deep Learning in NLP." Cornell University, arXiv:1906.02243 (5 June). https://arxiv.org/abs/1906.02243.

US SIF Foundation. 2020. "2020 Report on US Sustainable and Impact Investing Trends." www.ussif.org/files/Trends/2020_Trends_Highlights_OnePager.pdf.

Wu, Kai. 2021. "Measuring Culture." Sparkline Capital (24 August). www.sparklinecapital.com/post/measuring-culture.

Yin, Wenpeng, Jamaal Hay, and Dan Roth. 2019. "Benchmarking Zero-Shot Text Classification: Datasets, Evaluation and Entailment Approach." Cornell University, arXiv:1909.00161 (31 August). https://arxiv.org/abs/1909.00161.