THE NEXT GENERATION ALPHA MODEL:

SEEDING INTELLIGENCE WITH THE GROUND TRUTH

VIRTUS SYSTEMATIC

THE SEEDS FOR INTELLIGENCE

Our white paper, *The Next Generation Alpha Model: Introducing Natural Language Processing*, introduced the application of artificial intelligence ("AI")-powered natural language processing ("NLP") in asset management. In this follow-up paper, we highlight how we have established a differentiated AI process designed to add alpha by detecting the sentiment of sell-side equity research models.

Developing any AI process can be thought of as employing a two-step approach: an initial seeding of the intelligence or knowledge for an AI model, and then engaging in continued self-learning through an ever-growing dataset.

For the initial seeding process, the AI model starts as anything but "intelligent," but, over time, it morphs into something smarter as it sees more data and gains experience. Here we draw an analogy between the development of a child's natural intelligence over time and developing an artificial intelligence model with the help of Tommy, a curious toddler.

Teaching an NLP model is similar to how Tommy's parents would educate him. For example, Tommy sees the following pictures, believing both are croissants.

However, Tommy's loving parents will inform (provide initial data) him that the photo on the right is actually a cat and not a French pastry. As this example illustrates, adults serve a critical function for a child's development, providing a frame of reference (initial data) that guides learning.

KEY TAKEAWAYS

– The Virtus Systematic Team's proprietary AI model is built on a carefully crafted learning process, designed to optimize detecting sentiment in equity research reports.

– While it is widely claimed that large datasets enable AI, the true seeds for intelligence come from the ground truth.

– Efforts made by the Systematic Team to create a gold-standard ground truth:

- ■ **Expansive dataset.** Our AI model is rooted in a dataset of 1.5 million (and counting) sell-side equity research models and reports.
- ■ **Objective/data relationship.** We preserved our model's predictive power by avoiding overfitting against performance.
- ■ **Objective labeled data.** Our proprietary systematic approach accurately labels the ground truth for robust AI model learning.

ESTABLISHING THE GROUND TRUTH

Just as children need role models to guide their growth during their initial years, AI systems need a ground truth—large training datasets with accurately labeled data points aligned with the model's objective. Think of a ground truth as a massive set of flashcards, each of which has a question on one side and the answer on the other. This standard guides a model's machine learning endeavors.

However, not all ground truths are created equal. In our experience, there are three critical components to create a "gold standard" ground truth, each of which carries distinct challenges:

1. Expansive dataset
2. Objective/data relationship
3. Objective labeled data

1. Expansive Dataset

Our NLP model is an artificial neural network, modeled to mimic the human brain – a biological neural network. Like a newborn child, an NLP model starts with a blank slate, requiring stimuli (data), experiences, and enrichment to grow. This growth comes in the form of an expansive dataset. The size of the training dataset often determines the quality of an NLP model. If it is too small, the model will likely fall short of its true potential.

Going back to the analogy of a child's development, if Tommy's parents spend 30 minutes reading to him every day, we should expect him to have a strong command of language. Likewise, the greater the multitude of examples and data an NLP model is exposed to for its training, the higher the chances that it is able to produce better results.

In earlier stages of this project, our dataset was comprised of 100,000 sell-side equity research reports. At the time of writing, it exceeds 1.5 million (and counting) reports, as we continue building relationships with brokerage houses around the world. Training the model with this expanded dataset that spans over 12 years (2008–2019) helped it learn how an analyst's sentiment can sour in a recession and how macro industry-level news can influence an analyst's thoughts towards stocks. The more data the model is exposed to, the stronger its predictive power.

2. Objective/Data Relationship

An artificial neural network is an incredibly powerful computing system designed for complex learning and pattern recognition. It can be a double-edged sword, as we will touch on later. Data scientists train neural networks with a training dataset which the AI will use a basis for recognizing patterns.

After the training process is complete, data scientists evaluate the neural network's accuracy against a testing dataset, which is entirely new to the model. The training dataset and testing dataset are typically called in-sample and out-of-sample data, respectively. A model with proper training will demonstrate a high level of accuracy for both in- and out-of-sample datasets.

To achieve a highly accurate outcome, we first carefully consider the data's relationship with our model's ultimate objective. In other words, we want to minimize the noise between the data and objective. The quality of this relationship sets up the foundation for future model efficacy and is something that needs to be resolved before writing a single line of code.

If the objective for an AI model is to identify cats in pictures, then the training dataset should be pictures of cats and pictures of cat look-a-likes labeled as "not cats." Here, there is a clear relationship between the cat identification objective and the dataset of cat and non-cat images.

For our NLP model, the objective is to detect the sentiment of sell-side equity research models. The dataset consists of sell-side research reports, which naturally carry the sentiment of the analyst expressing their opinion of a stock. This clear relationship between our sentiment objective and the dataset provides a pathway for an NLP model to learn.

Since our ultimate goal is to forecast asset prices from the results of our sentiment model, it might be tempting to have the returns of assets as the objective of our sentiment analysis. However, assuming a relationship between the objective and the dataset can derail results. This is a common pitfall reported by many practitioners developing NLP algorithms. Returns across a broad array of assets are dictated by a host of factors, which may not be captured in the training dataset. When there is an inconsistent and tenuous relationship between the data and objective, neural networks often shoehorn a pattern where one does not exist.

In an asset management application, return-trained models generate excellent returns for the training dataset because they memorize all the noisy price return patterns against sentiment. However, they fail in out-of-sample testing because sentiment is one of many drivers of asset prices. This modeling error is known as overfitting, which is potentially amplified when machine learning is applied.

While AI is a powerful analytical tool that can be powerfully wrong, we believe AI can meaningfully contribute to alpha if applied correctly. In our white paper, *The Next Generation Alpha Model: Introducing Natural Language Processing*, we showed that the predictive power of the model could be used as an input in another model that forecasts asset returns.

Going back to our parenting analogy, if we want Tommy to learn to identify cats correctly, we can use pictures of cats to achieve the objective. After this exercise, we should not expect Tommy to accurately guess the cost differential between a typical mixed-breed cat and a purebred Scottish Fold. We would need a completely different lesson!

continued

3. Objective Labeled Data

The next requirement for a high-quality ground truth is accurately labeled data aligned with the model's objective. For our model, we seek to identify the positive or negative sentiment of sell-side equity research reports. In practice, this means that each research report in the training dataset has an accurate sentiment label.

Inaccurate sentiment labeling interferes with an NLP model's machine learning process. With a meaningful amount of mislabeled data, the resulting NLP model will have a distorted sense of positive and negative sentiment. Just like if Tommy was wrongly shown pictures of dogs and hamsters along with cats, and constantly reinforced that all the images are cats—we can expect a grown-up Tommy to think most household pets are cats.

While correct dataset labeling is critical for developing a gold-standard ground truth, labeling a dataset that consists of millions of reports represents a monumental challenge. Moreover, labeling and categorizing a broker report based on its sentiment is much more arduous and nuanced than classifying pictures of cats and non-cats.

We executed this by combining our global resources, leveraging a proprietary systematic approach to perform initial labeling. We then used investment professionals and finance students from a top local university to further manually label the data. An accurately labeled dataset sets the NLP model up for the best possible learning and optimization outcome.

PREVENTING ARTIFICIAL UNINTELLIGENCE

Media and business leaders often claim that data is the key to unlocking the transformative application of AI. While this infers organizations with the largest data repository have the greatest potential for success, we believe data only represents one of the many preconditions for success. Data is the rawest input in the artificial intelligence value chain and needs to be refined to a ground truth to be usable; the gap between data and ground truth can be significant.

To illustrate the contrast, it is like having two-year old Tommy being left alone in a room full of books versus having been read those same books by his parents. Between the former laissez-faire and latter guided-learning scenarios, there is a significant difference in the potential quality of learning.

While the concept of AI certainly carries the aura of cutting-edge computer science, intelligence is not guaranteed. Indeed, an artificial intelligence model and a naturally intelligent Tommy are both sponges, absorbing and learning from encountered experiences. Tommy's formative years will have an exceedingly strong influence on the type of adult he will become. Similarly, we are mindful of our AI model's formative moments. As we continue to build and refine the next-generation alpha model, our success rests on meticulously constructed ground truths.

continued

Kunal Ghosh
Senior Managing Director
Chief Investment Officer

Lu Yu, CFA, CIPM
Managing Director
Senior Portfolio Manager

Jie Wei, CFA
Director
Senior Portfolio Manager

Yang Zhang
Assistant Director
Data Scientist

The Virtus Systematic team of Virtus Investment Advisers, Inc. specializes in differentiated investment solutions, strategies, and outcomes across asset classes, regions, and securities. The Virtus Systematic team manages U.S. mutual funds for Virtus Investment Advisers, Inc. and other portfolios for Virtus Fund Advisers, LLC.

This report was originally published in October, 2019 while the Team was known as the Allianz Global Investors U.S. LLC ("AllianzGI") Systematic Investment Team. Effective July 25, 2022, AllianzGI's Systematic Investment Team joined Virtus Investment Partners, Inc. as the Virtus Systematic Team. The content has been edited from the original.

The commentary is the opinion of Virtus Systematic. This material has been prepared using sources of information generally believed to be reliable; however, its accuracy is not guaranteed. Opinions represented are subject to change and should not be considered investment advice or an offer of securities.

Past performance is no guarantee of future results.

All investments carry a certain degree of risk, including possible loss of principal.

4693 11-22

VIRTUS
SYSTEMATIC
virtus.com • 1-800-248-7971