Unlocking the Secrets in Semantics

What is Natural Language Processing or “NLP”?

Natural Language Processing or “NLP” is a branch of Artificial Intelligence, which aims to enable machines to be able to read, decipher, understand, and ultimately make sense of human language in a manner that is of value. NLP is increasingly automating operational processes ranging from the simple; answering a question from the internet, to the more complex; processing gigabytes of unstructured data, and generating terminologies, making implicit connections, and inferring that data’s context.

Today, NLP is the driving force behind some of the most commonly-used applications across our day-to-day lives:

- Language translation applications such as Google Translate
- Word Processors such as Microsoft Word and Grammarly that employ NLP to check the grammatical accuracy of text
- Interactive Voice Response “IVR” applications used in call centres to respond to certain users’ questions and requests
- Personal assistant applications such as OK Google, Siri, Cortana, and Alexa

The NLP community’s current focus is on exploring several key areas of research, including; semantic representation, machine translation, textual inference, and text summarization.

Certainly, the recent advancements in Machine Learning techniques have enabled data scientists to advance these techniques hand in hand. Data is being generated and captured at an exponentially increasing rate, and NLP is an important tool in our box to enable us to better understand what is happening across global markets.

What are the challenges of using NLP in Finance? (data)

Specific to what we do (systematic investing), traditional market and factor data are typically structured in numerical terms and are relatively simple to use within the machine or deep learning models. However, despite the abundance of rich textual data taken from financial news, earnings reports, and transcripts and their correlation to markets, currently, quantitative managers rarely exploit this text data for the following reasons:

- Firstly, raw textual data is represented by its categorical and symbolic features, which presents a problem for quantitative models. However, one key NLP technique, which could help overcome this issue, is language representation (i.e. text embedding). This technique transforms text symbols into numerically digestible high-dimensional (i.e. several hundred or thousands) dense vectors, while importantly still preserving semantic closeness.
  Traditional (count-based) feature engineering strategies for textual data involve models belonging to a family of models popularly known as the “Bag of Words” (insert definition) model. This includes term frequencies, TF-IDF (term frequency-inverse

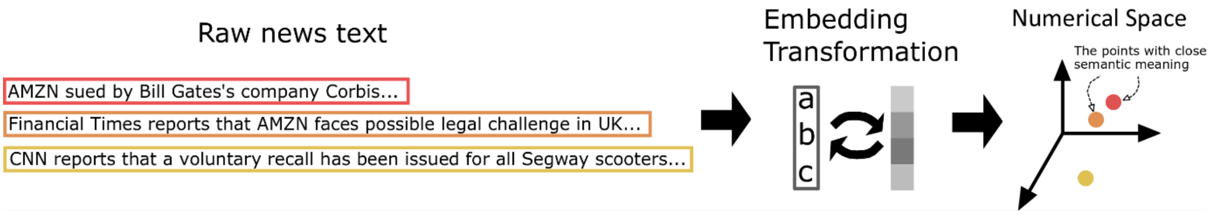
document frequency), N-grams and so on. While they are effective methods for extracting features from text, due to the inherent nature of the model being just a bag of unstructured words, we lose additional information like the semantics, structure, sequence, and context around associated words in each text document. For us, this represents an opportunity to explore more sophisticated models that can capture this information and provide us with features that are vector representations of words, popularly known as embeddings.

- Secondly, financial data comes from diverse sources, spanning financial news, earnings reports, and transcripts, etc. The consequence is the variety of data formats and structures. Being able to adapt the text embedding to different data sources in order to capture a variety of meaningful information, is vital.

- Thirdly, financial text data is sparse. For instance, financial news moves in-parallel with real-world events (i.e. at sporadic times), while earnings reports are routinely released monthly or quarterly. This is in sharp contrast to the structured and well-formatted market and factor data typically consumed by our quantitative models. Therefore, it's crucial to develop a systematic approach to encoding sparse text data for quantitative models.

Being able to leverage NLP across real-time voice transcriptions and chat can provide additional data that can be integrated into our deep learning models. In terms of finance, NLP can become a powerful tool for asset managers to discover actionable insight from the realms of unstructured data that is produced throughout markets.

Ultimately, we believe that the implications of NLP are profound and extremely positive for augmenting our current data sets for us to better-capture signals which can help us understand the markets in which we invest.


Raw news text
AMZN sued by Bill Gates's company Corbis...
Financial Times reports that AMZN faces possible legal challenge in UK...
CNN reports that a voluntary recall has been issued for all Segway scooters...
Embedding Transformation
a
b
c
Numerical Space
The points with close semantic meaning

Source: RAM Active Investments

Illustration of text embedding. Raw news text is transformed into dense vectors in the numerical space. The different colours represent the correspondence between text and vectors (i.e. points) in the space. Texts with semantic closeness are mapped to points close in the space.

What are the challenges of using NLP in Finance? (models)

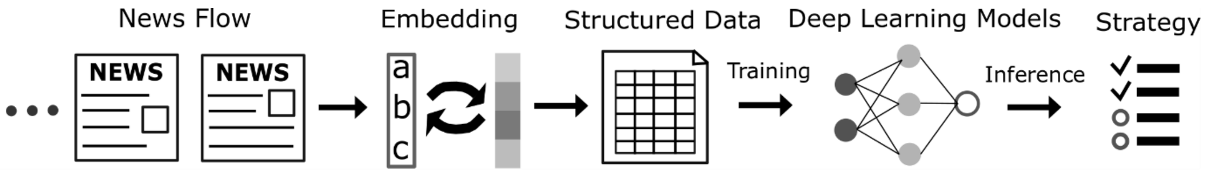
Aside from the challenges we discussed in terms of data, designing a suitable deep learning or quantitative model in addition to financial text is an extremely complex task. This is in-part owing to financial text data not playing a standalone role, but instead having a close relationship with market data, i.e. we know that various fundamental factors underpin market behaviour. At RAM AI, we have developed a deep learning model, capable of consuming both factor and text data to help capture their interactions and subsequently their effects on the wider market. As we know, the financial text is relatively sparse, together with its high dimensionality and the associated noise, a deep learning model would need to be training in an extremely robust manner to identify and capture genuine patterns. Diving deeper into the challenges of using NLP effectively, the very nature of the differing regions, markets and sectors requires an adaptable and robust quantitative model. From a machine learning perspective, this can be achieved by “transfer learning”.

How has RAM AI adapted and adopted NLP into their research process?

At RAM AI, our current NLP and Deep Learning efforts are twofold.

First, we applied state-of-the-art text mining and NLP techniques to extract information from finance text (for example; news, transcripts, earnings reports, etc.) and transform them into quantitative “model friendly” features. Then, in order to further boost our existing market prediction models, we are developing specialised deep learning architectures and learning procedures across both textual features and fundamental factors. This will enable us to exploit the synergy of fundamental and text data for a more accurate prediction, and ultimately alpha generation. Moreover, our NLP and deep learning pipeline is generic and highly flexible, with the ability to adapt to different market segments depending on our interest.

Second, to identify in real-time certain events or aspects of interest in the market by information extraction from diverse financial data. For instance, by plugging our in-house NLP pipeline into financial news flow, we can pinpoint company-specific ESG events and then inform clients in a timely manner.



Source: RAM Active Investments

A simple example of integrating news flow into strategy design.

What is NLP's value-added?

It is widely accepted that real-world events reflected within unstructured data, e.g. financial news, earning calls, transcripts, financial reports, social media, etc, have a certain relationship to markets. NLP enables us to integrate inputs from these unstructured and qualitative data sources into our quantitative models. These inputs, which are complementary to our existing quantitative/structured inputs from analysts' revisions, enrich the information set that our quantitative models consume.

Meanwhile, with the quantitative models enhanced by unstructured data, the subsequent strategy selection process can react to real-time events and capture potential investment opportunities in a more dynamic and timely manner. For instance, more conventional climate and ESG related data are at a low frequency and data providers would typically take days or weeks to react, while automatically analysing news flow helps us to identify the latest ESG related issues on companies and assess their wider impact.

What are the implications in the world of finance?

The implications for NLP and deep learning techniques for money managers are profound. If implemented correctly their use expands the horizon of data in terms of variety, volume, and velocity, where variety is the type of data, volume represents the amount of data automatically processed, and velocity means we can process the frequently arriving data, faster.

These techniques can help to further automate quantitative investing. For these unstructured data typically processed by analysts before, we are now able to seamlessly integrate them into quantitative models with less human involvement, the associated inherent biases and time delays.

For RAM this automated pipeline does not infer the replacement of humans by AI. On the contrary, it serves to highlights the importance of domain knowledge and expertise in this field. This is because we still rely on domain knowledge to tailor the models, and to guide the algorithms and models to focus on important aspects of large-scale financial data.

A brief history of NLP

Back in the late 1970s and early 80s, there was a revolution across artificial intelligence and machine learning. The general approach was to avoid complex decision trees of hard and fast rules and ultimately treat problems with Bayesian judgement.

Within the world of NLP, that meant not looking to understand the meaning behind each word, but to assume there was a hidden underlying meaning that language displayed hints about. The algorithm considers a range of meanings and tries to zero in on the most likely ones as more and more words are examined.

1906–1911, The Swiss linguistics professor Ferdinand de Saussure, changed the “Language as science” to “Language as systems”.

1950, The scientist Alan Turing further developed that idea in his research paper “Computing Machinery and Intelligence”, which became a milestone in NLP research.

1958 - 1964, grammar rules-based NLP research.

1966, Artificial Intelligence and Natural Language Processing (NLP) research were considered a dead end by many (though not all).

1980s, Machine learning began to be applied in NLP.

1997, Juergen Schmidhuber and his students from Dalle Molle Institute for Artificial Intelligence Research (Switzerland) introduced Long-Short-Term-Memory recurrent neural networks (LSTM), which were then widely used for voice and text processing.

2001, The first neural language model by Yoshio Bengio using a feed-forward neural network.

2008, Collobert and Weston spearheaded ideas of pre-training word embeddings in their award-winning paper “the Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning,”

2010, Recurrent neural networks used for language modelling.

2013 -2016, Dense vector representations of words for transfer learning, e.g. Word2Vec, GloVe, FastText.

2014, Convolutional neural networks applied to natural languages.

2014, Sequence-to-sequence learning for machine translation.

2015, Attention mechanism in sequence-to-sequence learning.

2017, Self-attention-based Transformer architecture achieving state-of-the-art for Neural Machine Translation.

2018 – 2019, Transformer based pretrained language model, e.g. BERT, RoBERTa, XLNet, ALBERT, etc.