MULTI-DIMENSIONAL ALPHA

August 27, 2019

NLP 5G EVOLUTION

Text Mining Global Corporate Filings

- **Introducing our Fifth Generation (5G) NLP Model.** Continuing our quest for alternative signals from unstructured textual data, in this research, we propose the Fifth Generation (5G) of our NLP/ML modeling framework. The focus of our 5G model suite is annual and interim corporate regulatory filings, sourced from the SEC's EDGAR database for US companies and an alternative data vendor Mergent for international firms.
- **Cloud-Based Technology Infrastructure and Innovative NLP/ML Algorithms.** Computing on textual documents is far more difficult and much slower than numerical data. In our 5G system, we have revamped our technology infrastructure, by taking full advantage of cloud computing, noSQL storage, and GPU parallel processing. On the modeling side, we develop our own in-house NLP library which goes beyond conventional sentiment analysis. Furthermore, we incorporate psychology research, behavioral finance, and corporate governance in our models. More importantly, we adopt the latest advance in deep learning and boosting. Our CNN (Convolutional Neural Networks) powered models extract additional information from company filings.
- **Multilingual Global Text Mining of Annual and Interim Filings.** The international annual and interim corporate filings from Mergent are in PDF form, which presents additional challenges. There are considerable other obstacles in processing non-US corporate filings. There is no standardization across countries. Furthermore, filings are in 21 different languages. While the vast majority of existing NLP algorithms are designed to process English language, we have developed multilingual dictionaries, invented language specific tokenizers, and re-adapted the CNN algorithm. The three major Asian languages – Japanese, Chinese, and Korean pose the largest challenge in custom NLP analysis. In the end, our models cover more than 2,000 public companies in Europe, Asia, LATAM, and Canada, in addition to the full coverage of Russell 3000 in the US.
- **Introducing the GINA Model.** Lastly, we combine our NLP signals, machine learning driven sentiment analysis (via Elastic Net and xgBoost), along with our deep learning algorithm (CNN) together in the new GINA (Global Intelligence NLP Alpha), which supersedes the previous SPEC. The GINA model delivers strong and uncorrelated performance in the US markets, with a Sharpe ratio of 1.0x. The GINA model covers over 2,000 stocks in Europe, Asia (including Japan and China), LATAM, and Canada, with investment horizon beyond a year.

English	Swedish	Chinese	Korean	French	German	Japanese	Russian	Thai.Lao	Portuguese	Spanish
achieve	uppnå	实现	이루다	atteindre	leisten	達成する	достигать	บรรลุ	conquistar	lograr
bad	dålig	坏	나쁜	mal	Schlecht	悪い	Плохо	ไม่ดี	mau	malo
challenging	utmanande	具有挑战性的	도전적인	difficile	herausfordernd	挑戦	испытывающий	การท้าทาย	desafiador	desafiante
denial	avslag	否认	부정	le déni	Verweigerung	拒否	отказ	การปฏิเสธ	negação	negación
efficacy	effektivitet	功效	효능	efficacité	Wirksamkeit	効能	эффективность	ประสิทธิภาพ	eficácia	eficacia
excellent	excellent	优秀	우수한	excellent	Ausgezeichnet	優れた	отлично	ยอดเยี่ยม	excelente	excelente
excited	upphetsad	兴奋	흥분한	excité	aufgeregt	興奮した	в восторге	ตื่นเต้น	animado	emocionado
good	Bra	好	좋은	bien	gut	良い	хорошо	ดี	Boa	bueno
great	bra	大	큰	génial	groß	すばらしいです	Великий	ยิ่งใหญ่	ótimo	estupendo
insecure	osäker	不安全	불안정한	peu sûr	unsicher	安全でない	небезопасный	ไม่ปลอดภัย	inseguro	inseguro

Gaurav Rohal, CFA
GRohal@wolferesearch.com

Yin Luo, CFA, CPA
YLuo@wolferesearch.com

Sheng Wang
SWang@wolferesearch.com

Javed Jussa
JJussa@wolferesearch.com

Miguel Alvarez
MAlvarez@wolferesearch.com

Jason Zhong, PhD
JZhong@wolferesearch.com

Zhao Jin
ZJin@wolferesearch.com

QES Desk: 1.646.582.9230
Luo.QES@wolferesearch.com

Table of Contents

.....1

A Letter to Our readers.....3

QES Text Mining Infrastructure for US Companies 5

The SEC EDGAR Corporate Filing Database 5
QES Text Mining Infrastructure 6

A New Generation of NLP and Machine Learning Algorithms 13

Language Similarity Signal..... 13
Sentiment Analysis and Behavioral Finance 21
The Next Frontier – CNN (Convolutional Neural Networks) 26
CNN Introduction 27
Introducing the GINA (US) Model..... 30

Text Mining International Annual Reports and Interim Filings 34

Tokenizing Non-English Languages..... 37
Language Similarity Signal..... 37
Sentiment Analysis and Behavioral Finance 40
CNN – International..... 45
GINA (International) Model 46

An Uncorrelated Source of Alpha.....49

Online Tools and Data Feeds.....51

Bibliography.....55

Disclosure Section.....58

A LETTER TO OUR READERS

As highlighted in *The Future of Active Management* (see Luo, et al [2019]), generating alpha has become increasingly difficult everywhere across global markets over the years. The excess returns from traditional factors seem to be converging to zero, with increasing confidence. For well over a decade, we have been proponents of utilizing alternative data and machine learning techniques in active investing. In particular, the large volume of unstructured textual and imagery data provides great opportunities of untapped alpha. Due to the nature of the low signal-to-noise ratio, extracting insights from unstructured text remains a significant challenge. In this research, we are introducing the fifth generation of our NLP (Natural Language Processing) and ML (Machine Learning) model – the GINA (Global Intelligence NLP Alpha).

The first generation of NLP research started in mid-2000s. Seminal papers by Tetlock [2007, 2008] and Loughran and McDonald [2011] laid the foundation of textual sentiment analysis in finance. Data vendors such as RavenPack and Thomson Reuters (and now Refinitiv) commercialize news sentiment analysis and make it easily accessible to investors for a modest cost. As off-the-shelf news sentiment data becomes more and more widely used in the investment community, its alpha shrinks and its investment horizon shortens.

The next generation of text mining takes the preprocessed news and social media sentiment data as one input source, and then combines that with behavioral finance and machine learning techniques. For example, in *Beyond Fake News* (see Rohal, et al [2019]), we found that depending on the nature of the news, investors could either under- or over-react to news, which leads to different post-event drift. Combining news sentiment with corporate events using machine learning techniques, our NICE (News with Insightful Categorical Events) model delivers much stronger alpha with longer investment horizon than traditional sentiment factors from data vendors directly.

In the third generation of NLP algorithms, we introduced tools that analyze corporate regulatory filings in the SEC's EDGAR database (see *Text Mining Unstructured Corporate Filing Data*, Rohal, et al [2017]). In particular, we found a finance-specific sentiment dictionary can help us extract signals unique to company filings. Furthermore, changes in the language structures are highly predictive of future firm downside risk. In the end, we introduced our SPEC (Systematic Profiling EDGAR Composite) model that generates investment insights from 10-K/10-Q documents.

Most recently, we launched our SMEC (Systematic Mining of Earnings Calls) model (see *Tone at the Top? Quantifying Management Presentation*, Rohal, et al [2018]). The SMEC model further leverages machine learning (e.g., Latent Dirichlet Allocation) and psychology research to analyze thousands of management presentations and conference calls every day to predict the returns of more than 2,000 US companies and nearly 1,200 stocks outside of the US.

Although most of our first four generations of NLP models are global in nature – covering both US and international companies, the underlying NLP algorithms can only process English language documents.

This paper showcases our latest research in NLP/ML, i.e., the fifth generation (5G). For US companies, we revisit corporate filing data from the EDGAR database. The GINA (Global Intelligence NLP Alpha) model – an extension of the SPEC – can meaningfully improve performance by incorporating more advanced NLP and machine learning techniques (e.g., deep learning). More importantly, we extend our coverage from US to global. There are multiple complications when we analyze non-US companies.

First, there is no consistent filing format across countries. We source the annual and interim corporate filings from an alternative data vendor called Mergent. A much greater obstacle is to design proper NLP algorithms for non-English languages. As the vast majority of existing research centers around the English language, we make a heavy investment to develop our own non-English language NLP library. We use multilingual lexicons to categorize the text contained in corporate filings into sentiment – positive or negative (bullish or bearish). We look at the language complexity of textual content with part-of-speech tagging. As we previously highlighted, when firms make an active decision to significantly alter the wording and structure of language embedded in their regulatory filings, these conscious adjustments have considerable implications for future firm behavior. We use Word2Vec algorithm and non-English language tokenizers to measure these subtle language changes for international corporate filings.

In the end, we introduce our GINA stock selection model, leveraging sophisticated web scraping, cloud computing, and NLP/ML algorithms. The GINA model delivers consistent performance, with low turnover, slow decay, and is uncorrelated to traditional factors and fundamental investment styles.

How Clients can use our Research

For fundamental investors, the GINA model and selected factors are accessible from our online portal. Our NLP-powered GINA model has extremely low signal decay and is capable of extracting relevant information from long and tedious text of annual and interim filings.

For quantitative investors, the complete backtesting results are available upon request, which should assist managers in their own research. More importantly, we offer daily data feeds that investors can plug into their own investment process directly. Given the heavy usage of alternative data and machine learning techniques, we expect our models to be materially different from traditional fundamental factors; therefore, should provide significant diversification benefit.

Regards,

Yin, Gaurav, and the QES team

QES TEXT MINING INFRASTRUCTURE FOR US COMPANIES

Since the launch of our SPEC (Systematic Profiling EDGAR Composite) model in 2017 (see *Text Mining Unstructured Corporate Filing Data*, Rohal, et al [2017]), we have significantly improved our web scraping, NLP and machine learning algorithms. In this section, we introduce a new and improved version of the SPEC model for US companies. In the next section, we will discuss how to develop a NLP model for international companies using annual reports and interim filings in multiple different languages. The combined US and international text mining suite of algorithms is called the GINA (Global Intelligence NLP Alpha) model.

Early stage academic research mostly focused on textual sentiment analysis using pre-defined dictionaries. In recent years, we have seen tremendous progress in both academia and in the industry, with new NLP techniques such as readability, part-of-speech tagging, language specific tokenization, word vectorization, topic modelling, etc. A detailed review of these techniques and how to apply them in investment management can be found in *Tone at the Top? Quantifying Management Presentation* (see Rohal, et al [2018]).

As shown in Figure 1, there are hundreds of 10-K (annual) and 10-Q (quarterly) filings by US companies every day, especially during earnings season. Regulatory requirements also push the length of the 10-K/10-Q documents to become longer and longer. Few investors would have the time to read the voluminous corporate filings – not to mention being able to understand, digest, and generate investable decisions from these documents. Regulatory filings are typically written by accountants, lawyers, and professional editors, with standard language structure. The length and complexity of these filings make them ideal for text mining algorithms. In this paper, we show how unstructured textual data, behavioral finance, NLP, and sophisticated machine learning techniques can be brought together to predict future company performance.

THE SEC EDGAR CORPORATE FILING DATABASE

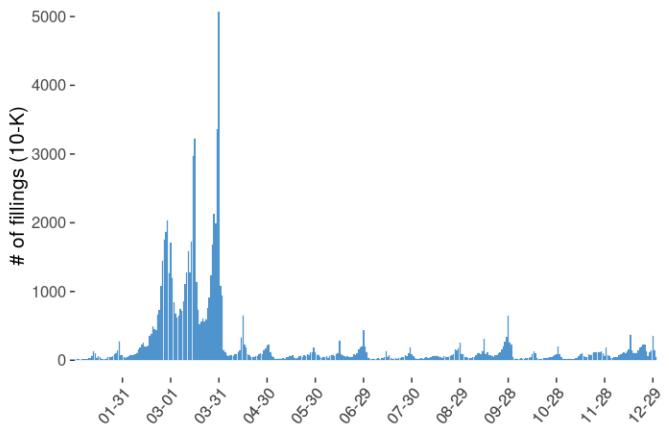
For US publicly listed companies[1], we use the EDGAR (the Electronic Data Gathering, Analysis, and Retrieval) system as our source of corporate filings. The EDGAR database is developed and maintained by the SEC (Securities and Exchange Commission) to disseminate business relevant information to the public. The SEC provides a full text version of filings called complete submission text file as well as a browser friendly version. The browser friendly version lists the core 10-K filing in a HTML format, including eight exhibits, two graphics files, and six XBRL documents. For the scope of this research, we focus only on the textual content of the 10-K/10-Q. Detailed discussion on the organization of the EDGAR database, trends on financial data reporting and web scraping is covered in our previous publication (see *Text Mining Unstructured Corporate Filing Data*, Rohal, et al [2017]).

As shown in Figure 1(A), the vast majority of 10-K filings appear in the EDGAR database in February and March. Most companies have December 31st fiscal year end. The SEC sets a deadline of 90 days after fiscal year end for annual reporting. As a result, the 10-K filings peak right before the end of March. On the other hand, 10-Q quarterly filings spike in the second month post each quarter end (e.g., February, May, August, November), coinciding the SEC's 45-day deadline for interim reporting.

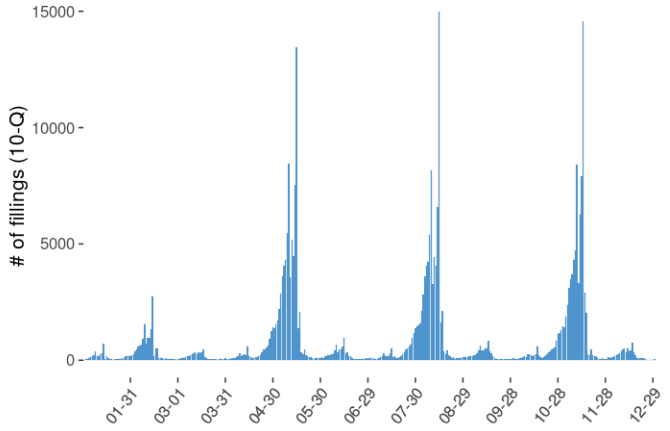
[1] In the next section, we will discuss how to use the Mergent database to extract annual reports and interim filings for non-US public listed companies.

Figure 1 Seasonality of the 10-K and 10-Q EDGAR Filings, Russell 3000

A) # of 10-K Filings



B) # of 10-Q Filings



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

QES TEXT MINING INFRASTRUCTURE

The first challenge in building a production quality NLP/ML model is system design. The textual data from the EDGAR database is highly unstructured. NLP/ML algorithms are computationally intensive and notoriously slow. It is unimaginable to manually download the millions of text files from the EDGAR website. Similarly, it is hopeless to conduct NLP/ML on even the most powerful PC on the planet.

Figure 2 summarizes our technology infrastructure setup for US companies. In the next section, we will demonstrate how to process international corporate filings, which face further obstacles. Our web crawler queries the EDGAR website, download, and store all corporate filing documents into a distributed file system. The web crawler also processes EDGAR Index master files to extract correct corporate entities and merge with our main data warehouse (called L-Quant, see The Big and Small Sides of Big Data, Luo, et al [2017] for details).

Each filing is then parsed through an in-house XML parser, which extracts text from each tag in the filing, maps item keys and description based on the predefined EDGAR filing structure and performs sanity checks using HTML anchor tags in the document. Our fault-tolerant computing system handles the highly complex and often error-prone text files, by striking a balance between minor deviations from the standard format versus true data errors.

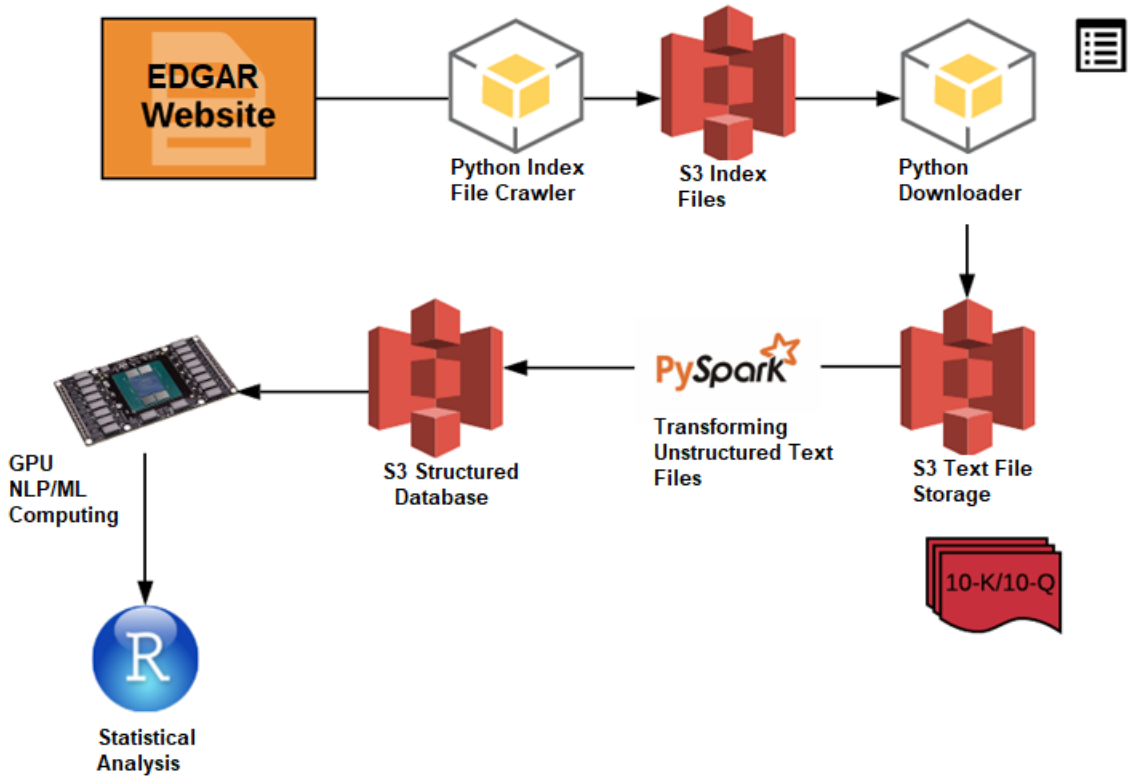
We use the AWS (Amazon Web Services) cloud computing platform to store, process and analyze textual information. For efficient storage and retrieval, we adopt the S3 (Simple Storage Service) – a distributed object storage service with low latency and high throughput supplied by AWS.

As explained *Text Mining Unstructured Corporate Filing Data* (see Rohal, et al [2017]), the raw text files in the EDGAR database are highly unstructured. Before our NLP/ML engine can be deployed, we need to pre-process and transform the unstructured data into well-defined sections. To optimally process the enormous volume of data, we take advantage of the PySpark programming – a collaboration of Apache Spark and Python. Apache Spark is an open-source cluster-computing framework, built for speed, ease

of use, and streaming analytics, while Python is a general-purpose, high-level programming language that is becoming the de facto language of choice for data scientists. A conventional setup on a local machine might be a good fit for academic researchers, where real-time analytics is not generally required, but would be hopeless for practitioners. The PySpark framework gives us the ability to conduct parallel computing with multiple nodes (machines and instances). We also make heavy usage of the EMR (Elastic Map Reduce) service from AWS. The EMR can process terabytes of data across a Hadoop cluster of virtual servers on Amazon EC2 (Elastic Compute Cloud) and S3 distributed data storage.

The core of our NLP engine is written in the Python programming language, fully developed by our team over the past decade. Our first generation of NLP engine was written exclusively in the R programming language. In recent years, we switched from R to Python, to leverage Python's exhaustive set of NLP/ML libraries. To further accelerate our text processing, we exploit GPU (Graphical Processing Unit) computing. A GPU is a specialized microprocessor, optimized to display graphics and images. While a typical CPU (Central Processing Unit) consists of four to eight cores, GPUs have hundreds or even thousands of cores. Although GPUs are designed primarily for special-purpose graph manipulation, we can tune them to perform mathematical operations. Leveraging GPU computing reduces the run time of our NLP/ML models from days to minutes.

Figure 2 QES NLP/ML Computing Framework



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

A New Fault-Tolerant Scraper

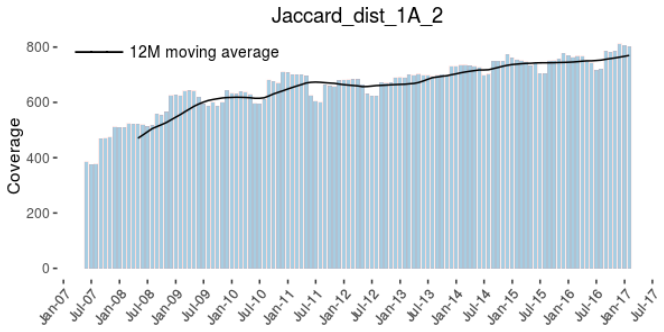
Since the EDGAR database contains millions of filings submitted by thousands of companies, the data is highly unstructured with many formatting errors and inconsistencies (see Stümpert [2008] and O´Riain [2012]), especially true in early years when the use of markup languages (e.g., HTML) was not as common (see Loughran and McDonald [2014]). Lack of fully tagged items, formatting errors and other inconsistencies lead to difficulties in accurately identifying and parsing text into structured sections.

Web scraping and transforming unstructured files into structured data require a balance between accuracy and reliability. Human errors, inconsistencies, and outliers are commonplace among websites, web-based document systems, and textual information. Overly emphasizing accuracy makes the algorithm too restrictive and filters out a significant chunk of information that does not fit within its constraints. On the other hand, relaxing the error bounds too much, although increasing data coverage, may end up incorporating erroneous information. There are no hard and fast rules on where to draw the line. It took us a considerable amount of time and effort over the past few years to build a proper fault-tolerant web scraper, specifically designed to process corporate filings on the EDGAR database. Compared to the first generation of the SPEC model, both coverage and data quality have improved substantially with the new scraper.

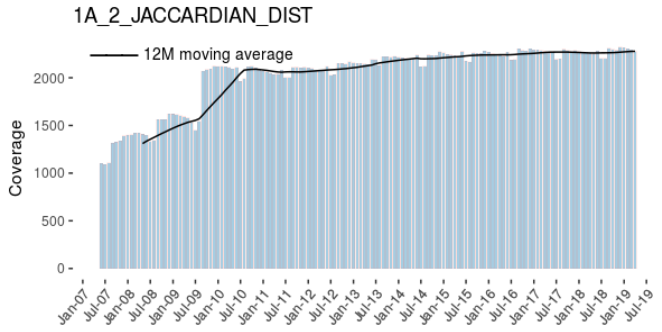
For example, in the SPEC model, we were only able to collect the “Risk Factors” section in the 10-Q filings for about 700-800 companies in the Russell 3000 universe – a coverage of 25% (see Figure 3A). With our new fault-tolerant system, the coverage is almost tripled to almost 2,400 companies (see Figure 3B). As highlighted in Text Mining Unstructured Corporate Filing Data (see Rohal, et al [2017]), the “Risk Factors” section provides critical insights on companies’ future performance and risk profile. Similarly, we also observe material enhancement in other sections, e.g, the “FSS” or “Financial Statements and Supplementary Data” (see Figure 3C and D).

Figure 3 Coverage Comparison, GINA(US) versus SPEC (10-Q Filings)

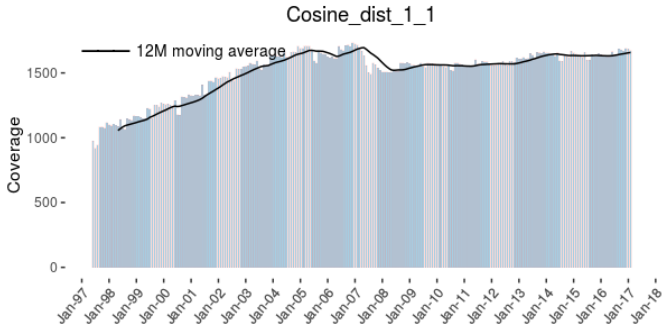
A) SPEC, Risk Factors Section



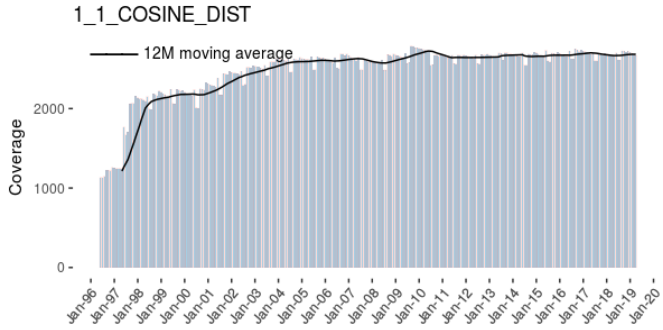
B) GINA, Risk Factors Section



C) SPEC, FSS



D) GINA, FSS



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo' s QES

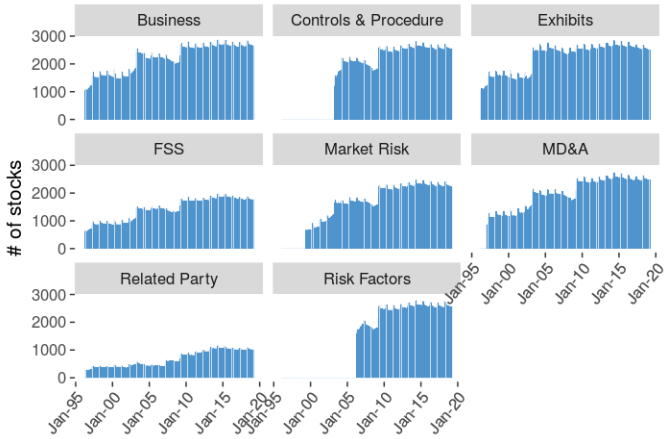
In addition to the "Risk Factors" and "FSS" sections, as shown in Figure 4, we have close to 90% coverage for the broad Russell 3000 universe in the US, for all eight major 10-K/10-Q sections – "Business", "Controls & Procedure", "Exhibits", "FSS", "Market Risk", "MD&A (Management's Discussion and Analysis)", "Related Party", and "Risk Factor". Our database starts from 1996 for most sections. The coverage for the "Risk Factor" section starts later, because it was a part of the "MD&A" until 2005.

In Figure 4, we can also see that coverage improves over time, especially after 2000. In the early years, we rely on the complete submission text files to extract each section. In recent years, companies have switched to a HTML format for their filings. The HTML tagging greatly improves the accuracy of our web scraper.

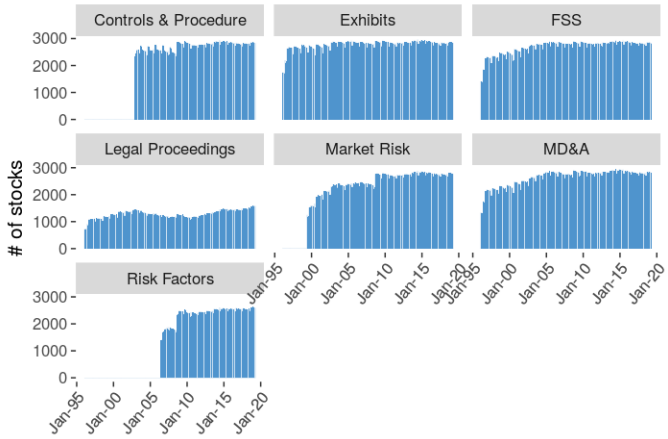
Even though the SEC sets clear disclosure requirements and reviews corporate filings to monitor compliance with the filing requirements, about 10% companies remain outliers, when it comes to adhering to the prescribed format.

Figure 4 US Public Company Coverage (Russell 3000 Universe), by Section

A) 10-K Filings



B) 10-Q Filings



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo' s QES

Poor Corporate Governance for the Non-Conformers?

Even after the wide adoption of HTML tagging after the last major overhaul to the "Risk Factors" section in 2005, we still have around 10% companies in the Russell 3000 index with missing parsed sections after 2006 (see Figure 5A). Due to the more complex nature of the 10-K filings, there are more companies with missing data in the 10-K than in the 10-Q filings.

There are multiple reasons that a company may decide not to adhere to the prescribed format by the SEC, e.g.,

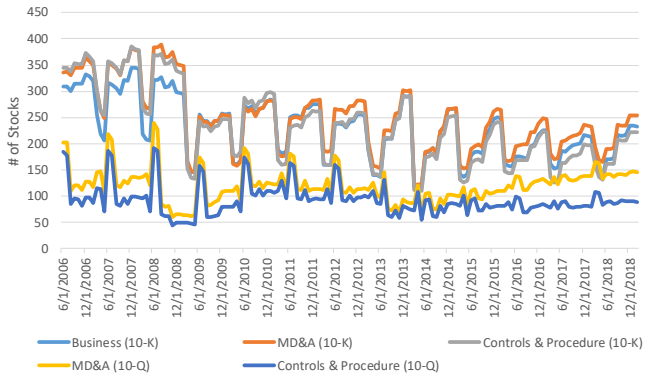
- Complex business model and organizational structure
- Weak corporate governance and control
- Lack of investment in technology infrastructure
- Intention to confuse and mislead investors
- Delayed filing

All above arguments indicate governance or operational weaknesses. Therefore, we hypothesize that the non-conformers potentially underperform the broad market. As expected, in the past 14 years, the stocks of those non-conforming companies underperform the Russell 3000 index by more than -4% per annum[2] (see Figure 5B). The strong seasonality observed in Figure 5(A) is possibly due to delayed filings, which also represents weaknesses in internal controls or accounting irregularities, as addressed *Quant CSI* (see Jussa, et al [2017]). Non-conforming companies tend to be smaller companies with higher shorting cost. The "limits to arbitrage" prevent investors from fully exploiting this market anomaly.

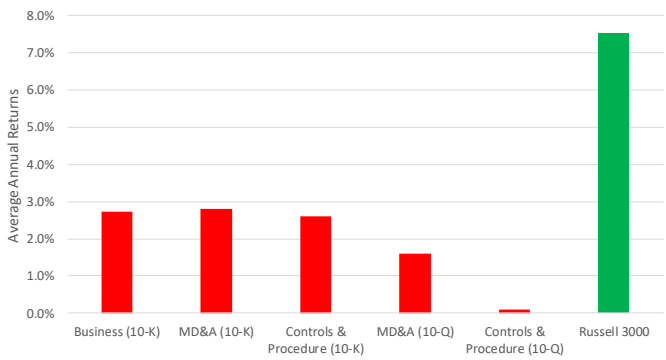
[2] Non-conforming stocks are equally weighted. The benchmark Russell 3000 index is also equally weighted.

Figure 5 Non-Conforming Companies to the Prescribed Format (Russell 3000)

A) Coverage



B) Non-Conformers vs the Russell 3000

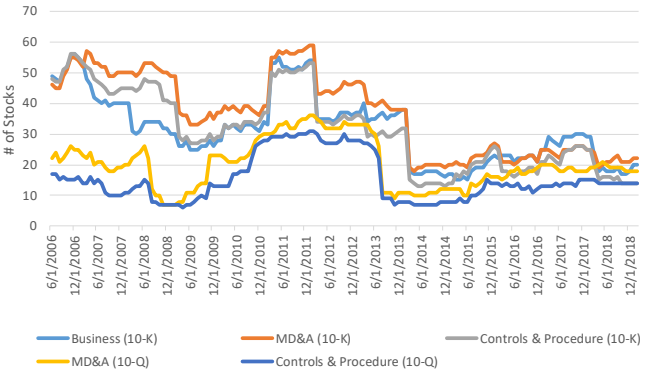


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

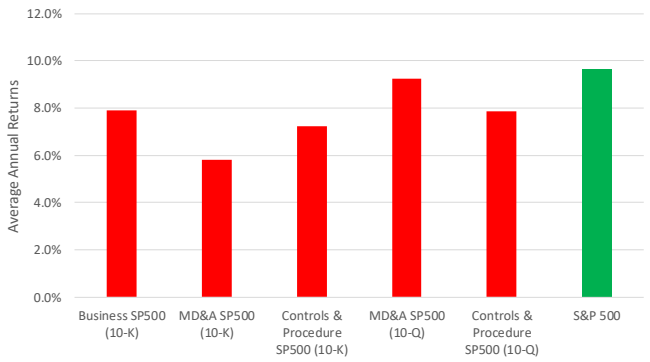
Next, we shift our focus to the non-conformers in the large-cap S&P 500 Index. There are far fewer non-conforming companies (around 20-50 stocks) in the large-cap universe (see Figure 6A). The strong seasonal patterns observed in the Russell 3000 universe (see Figure 5A) almost completely disappear in the universe of large-cap stocks. Interestingly, these non-conforming stocks still underperform the benchmark S&P 500 index, albeit the underperformance is not as precipitous as seen in the Russell 3000 universe. For both the Russell 3000 and the S&P 500 universes, missing "MD&A" section in the 10-K filings leads to the most striking underperformance. As one of the most critical sections, the "MD&A" provides the opportunity for management to elaborate firm performance and outline future strategies. Companies with missing "MD&A" section in the 10-K filing underperform the S&P 500 index by almost -4% per annum[3] (see Figure 6B).

Figure 6 Non-Conforming Companies to the Prescribed Format (S&P 500)

A) Coverage



B) Non-Conformers vs the S&P 500



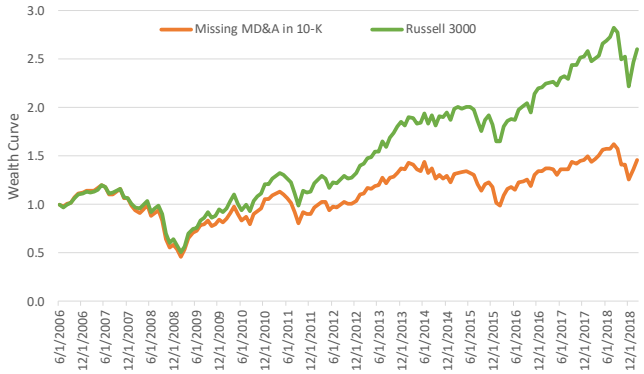
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

[3] Non-conforming stocks are equally weighted. The benchmark S&P 500 index is also equally weighted.

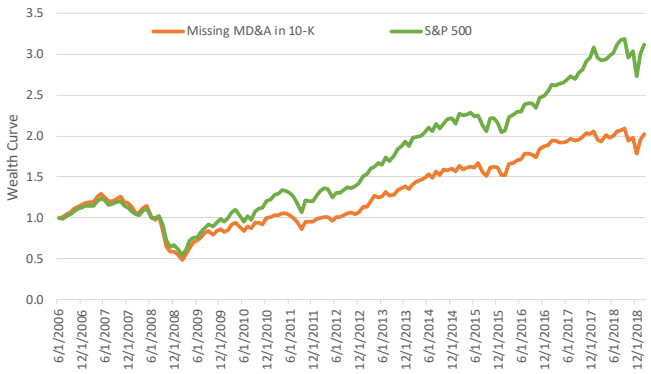
As shown in Figure 7, non-conforming stocks in both the Russell 3000 and the S&P 500 indices have consistently underperformed their respective benchmarks. We want to reiterate that the non-conformers are defined as those companies whose 10-K filings can't be parsed by the format prescribed by the SEC.

Figure 7 Cumulative Performance of Non-Conformers with Missing MD&A Section in the 10-K Filings

A) Russell 3000



B) S&P 500



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo' s QES

A NEW GENERATION OF NLP AND MACHINE LEARNING ALGORITHMS

In our SPEC model (see *Text Mining Unstructured Corporate Filing Data*, Rohal, et al [2017]), we applied the bag-of-words based sentiment analysis, language similarity, and behavioral finance to generate stock selection signals from the 10-K/10-Q corporate regulatory filings. For example, consistent with Cohen, et al [2010], we also found that firms with significant changes in their language structures (especially in the "Risk Factors" section) tend to materially underperform their peers and suffer considerably downside risk.

Since the launch of the SPEC model in 2017, we have substantially upgraded our NLP/ML libraries. For instance, in *Tone at the Top? Quantifying Management Presentation* (see Rohal, et al [2018]), we introduced readability test, part-of-speech tagging, VADER sentiment analysis, Word2Vec algorithm, topic modeling, LDA (Latent Dirichlet Allocation), and deep learning. These new algorithms help us to generate additional insights from management presentations and conference calls. In this section, we demonstrate how to apply these new techniques to analyzing 10-K/10-Q filings. The new and improved model is re-branded as GINA (Global Intelligence NLP Alpha).

LANGUAGE SIMILARITY SIGNAL

We start with the most interesting signal highlighted in the first-generation SPEC model – the language similarity score. In the SPEC model, we deploy two elementary document similarity measures, namely the Cosine similarity and Jaccard similarity.

Cosine Similarity

First, we need to compute two document term frequency vectors:

$$\begin{cases} D_1^{TermFreq} = [freqD_1^{t1}, freqD_1^{t2}, \dots\dots, freqD_1^{tn}] \\ D_2^{TermFreq} = [freqD_2^{t1}, freqD_2^{t2}, \dots\dots, freqD_2^{tn}] \end{cases}$$

Where,

D_1 and D_2 are two documents,

t_1 to t_n are the overlapping terms appearing in both documents, and

$D_1^{TermFreq}$ and $D_2^{TermFreq}$ are the sets of term frequencies in the two documents.

Equipped with the above two vectors, the Cosine similarity measure can be computed as:

$$Cosine\text{Similarity} = \frac{D_1^{TermFreq} * D_2^{TermFreq}}{||D_1^{TermFreq}|| * ||D_2^{TermFreq}||}$$

Essentially, Cosine similarity is the scalar product of the two document term frequency vectors divided by the product of their Euclidean norm.

Jaccard Similarity

On the other hand, Jaccard similarity is defined as:

$$Jaccard \text{ similarity} = \frac{D_1^{TermFreq} \cap D_2^{TermFreq}}{D_1^{TermFreq} \cup D_2^{TermFreq}}$$

Jaccard similarity is simply the intersection of the two document term frequency vectors divided by their unions.

The above two measures of similarity are highly correlated, with higher values indicating more similarity between two documents. In this section, we show how the Word2Vec algorithm can be used as another alternative similarity measure.

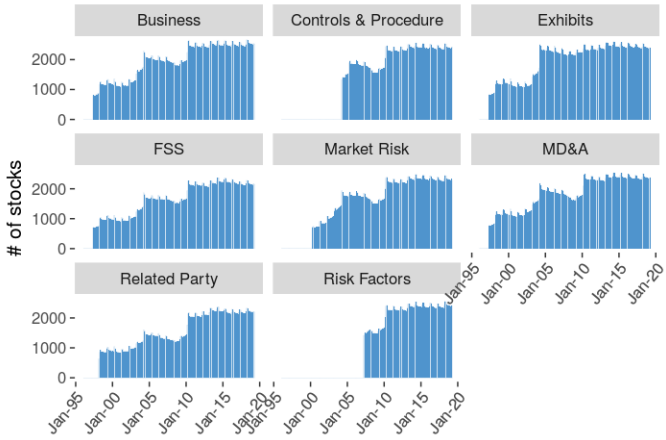
Word2Vec Similarity

In *Tone at the Top? Quantifying Management Presentation* (see Rohal, et al [2018]), we highlight that a vector representation of words or word embedding can overcome some of the impediments with traditional representation (e.g., latent semantic analysis). Given a collection of text documents, the Word2Vec algorithm (a two-layer neural networks) first constructs a vocabulary from the training text and then builds a vector representation of words. In this representation, words with similar meanings appear in closely related clusters. Word2Vec was published by Google in 2013. In our application, we use these vector representations of words instead of a simple term frequency count used in the SPEC model. Then, we can compute a Cosine similarity to statistically measure the correlation between two text documents. We name this factor as our Word2Vec Similarity measure.

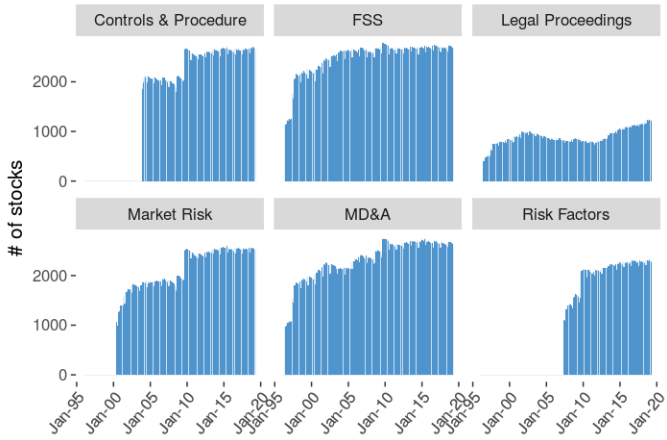
To compute the year-over-year changes in language structure, we do need filings from both current and previous years. As a result, coverage drops slightly (see Figure 8), compared to the total number of companies mapped in our database (see Figure 4).

Figure 8 Year-over-Year Word2Vec Similarity Measure Coverage, US

A) 10-K Filings



B) 10-Q Filings



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo' s QES

Language Similarity Factors

As shown in Figure 9, the performance of the Cosine and Word2Vector similarity factors are largely consistent. Both signals suggest that firms with larger changes in language structure – particularly the "MD&A", "Risk Factors", "Business", and "Control & Procedures" sections – tend to deliver lower subsequent stock returns. Payoff patterns in most sections, especially the "Business", "MD&A", and "Risk Factors" are mostly monotonic – higher levels of similarity (lower language change) corresponds

to higher future returns. Consistent with what we found in the SPEC model, the similarity signal based on the "Risk Factors" in the 10-Q filings delivers the strongest performance. The top 20% of companies with the least changes in their "Risk Factors" section have outperformed the bottom ones with the largest alterations by 5% per annum (see Figure 9).

Figure 9 Performance of Similarity Measure (YoY), 10-Q Filings

A) Cosine Similarity

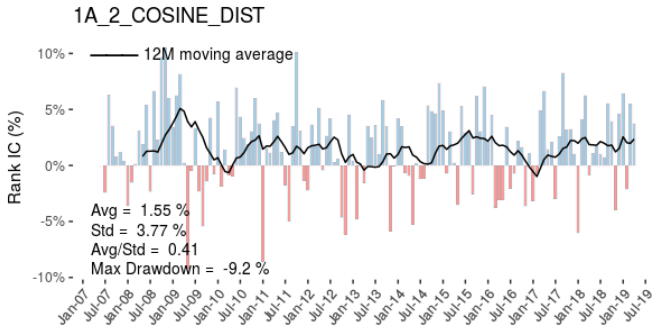
B) Word2Vec Similarity

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

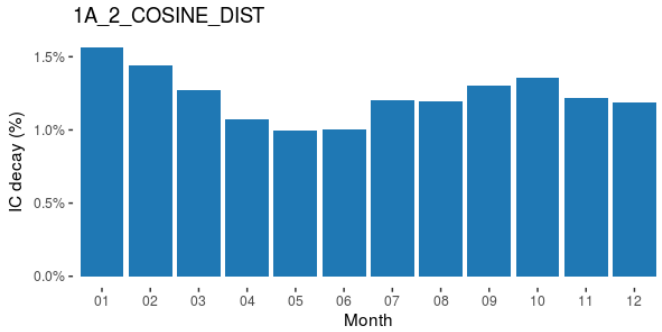
While the performance of traditional stock selection factors has deteriorated considerably in recent years, the language similarity signals continue to show strong predictive power (see Figure 10 A and C). While a monthly rank IC of ~1.5% may not be eye-popping, it is important to note their long investment horizon – both signals still show decent performance even after one year. Furthermore, as we will explain in a later section, most NLP-based models are almost completely uncorrelated to traditional stock-selection factors and fundamental investment styles.

Figure 10 Performance of Cosine Similarity Measure (YoY), 10-Q Filings

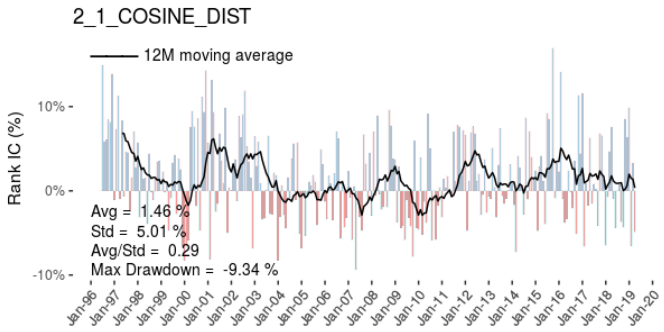
A) Rank IC ("Risk Factors" Section)



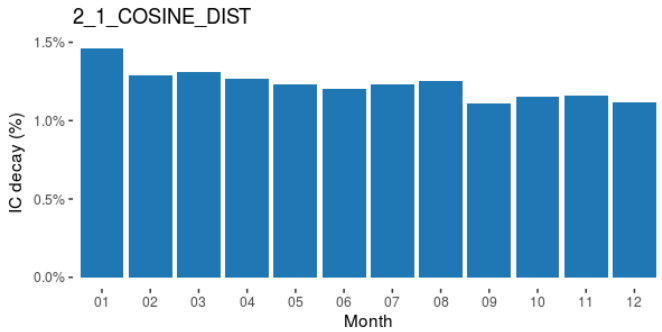
B) IC Decay ("Risk Factors" Section)



C) Rank IC ("MD&A" section)



D) IC Decay ("MD&A" section)



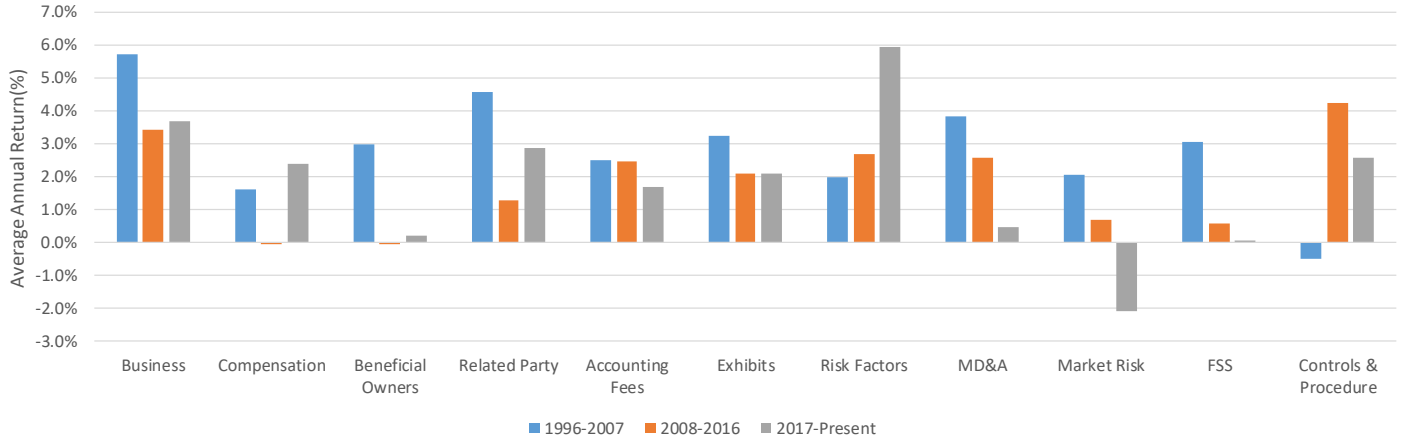
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

Figure 11 shows the average annual alpha of the long/short quintile portfolio based on each of the main sections in the 10-K filings, broken down into three periods – 1996-2007, 2008-2016, and 2017-Present[4]. Unlike the vast majority of traditional value/momentum/quality factors, the performance of our language similarity measures has been relatively consistent over the three periods.

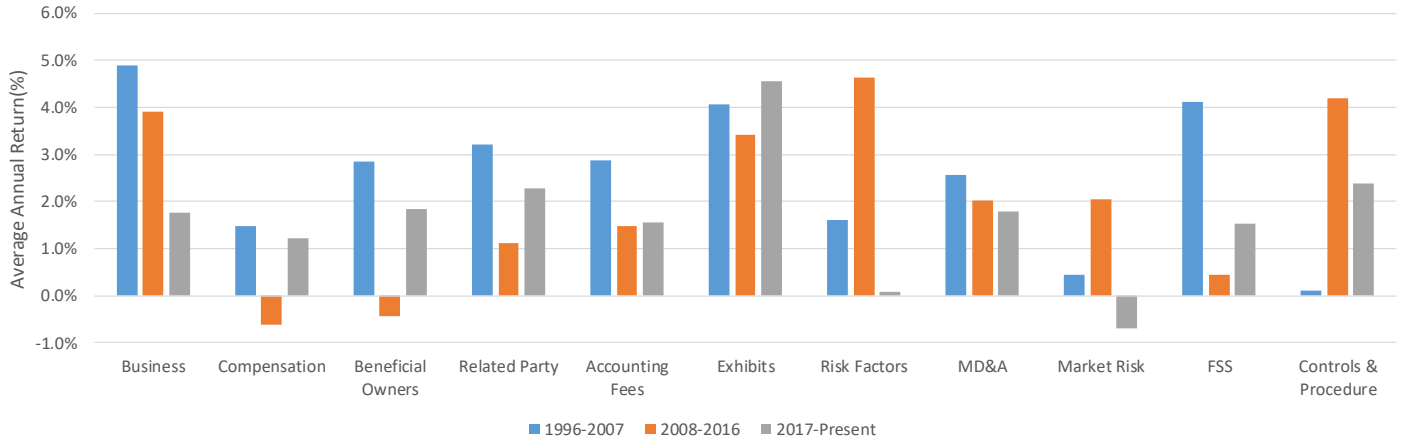
[4] The three periods correspond to the three important regimes of active investing. As detailed in *The Future of Active Management* (see Luo, et al [2019]), prior to the summer of 2007, even plain vanilla value/momentum factors had delivered exceptional performance. The summer of 2007 quant crisis, the 2008 Great Recession, and the subsequent March-May 2009 risk rally mark a new phase of quantitative investing. The dominance of unpredictable macro events (e.g., US-China trade war, Brexit) have caused further deterioration in factor performance.

Figure 11 Performance of Similarity Measures (10-K Filings)

A) Cosine Similarity



B) Word2Vec Similarity

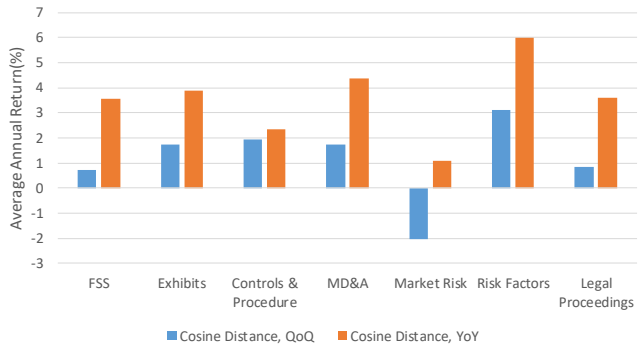


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

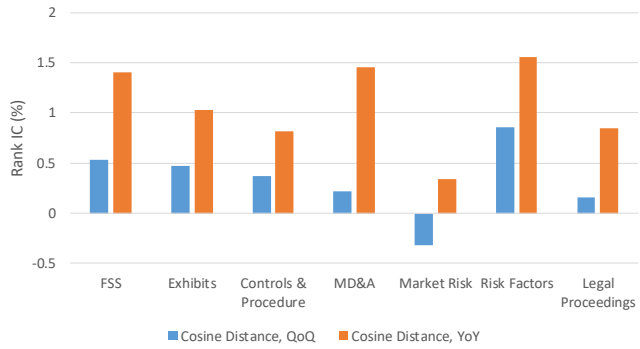
In addition, when we compare language similarity for 10-Q filings, we can compute either YoY (Year-Over-Year) or QoQ (Quarter-Over-Quarter) changes. YoY changes account for seasonality, while sequential QoQ computation reflects more timely information. As shown in Figure 12(A) and (B), while both measures are effective, YoY signals clearly dominate QoQ factors.

Figure 12 Performance of Similarity Measure, US 10-Q Filings (YoY vs QoQ Comparison)

A) Average Annual Returns



B) Rank IC



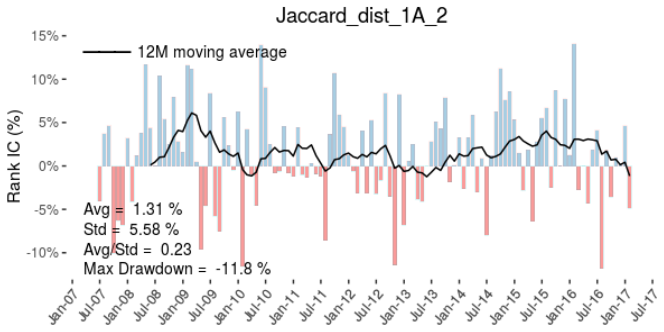
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

Performance Improvement

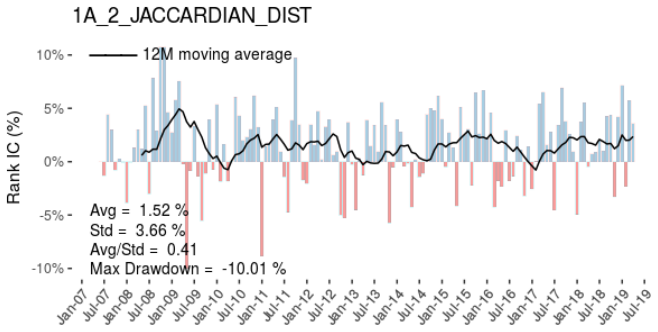
Lastly, to gauge the impact of our improved language similarity factors, we compare performance of the language similarity factors for the existing SPEC model and the new GINA. As shown in Figure 13, although the average rank IC's of the SPEC and GINA are roughly in line, the increase in breadth and coverage in the GINA model has reduced volatility and drawdown considerably.

Figure 13 Similarity Factor Performance Comparison, SPEC Versus GINA (10-Q Filings)

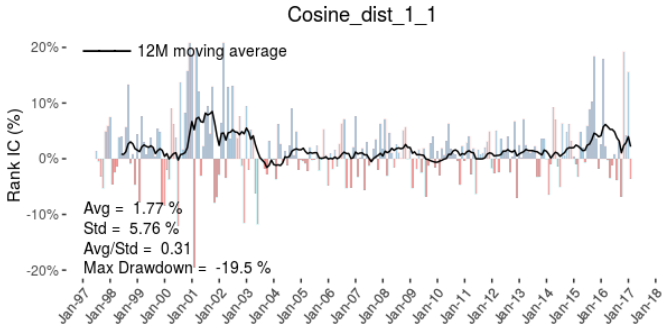
A) SPEC, "Risk Factors" Section



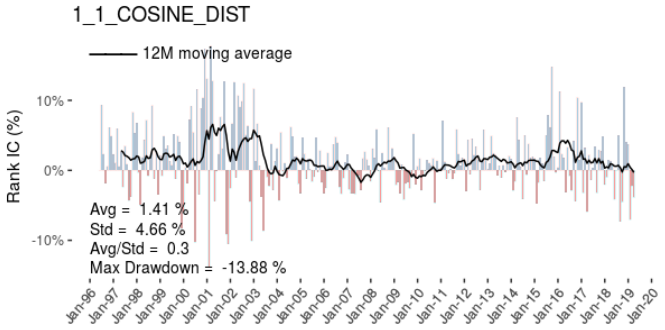
B) GINA, "Risk Factors" Section



C) SPEC, "FSS" Section



D) GINA, "FSS" Section



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo' s QES

A Case Study – Substantial Changes in the 10-Q Filings

Using Office Depot as an example, as shown in Figure 14(A), we can clearly see the new disclosures added in the "Controls and Procedures" section – the accounting of certain vendor program funds, the errors in timing of vendor program recognition, i.e., revenue recognition issues. Furthermore, in the "Legal Proceedings" section, the company added new disclosures that the SEC had started an inquiry regarding to the accounting error. Although it is obvious and easy to catch the changes on this occasion, in practice, it is extremely difficult to manually go through thousands of pages of filings for thousands of companies and still be able to identify new disclosures that have not been previously released. For NLP algorithms, however, it is a fairly straightforward task.

Figure 14 Office Depot, Q1/2007 versus Q1/2006

A) Q1/2007

Item 4. Controls and Procedures

Restatement

On October 29, 2007, Office Depot announced that its Audit Committee initiated an independent review principally focused on the accounting for certain vendor program funds. The Audit Committee, with the assistance of independent legal counsel and forensic accountants, assessed the timing of recognition of certain vendor program arrangements. The investigation revealed errors in timing of vendor program recognition and included evidence that some individuals within the company's merchandising organization failed to provide Office Depot's accounting staff with complete or accurate documentation of future purchase or performance conditions in certain vendor programs that would have otherwise required recognition of the related vendor funds to be deferred into future periods in accordance with the company's established practices.

As a result of the Audit Committee's review, on November 8, 2007, the Board of Directors of the company approved a restatement of the company's 2006 financial statements including corrections to amounts reported in the third and fourth quarters of 2006 and the interim financial statements for the first and second quarters of 2007, and the company is concurrently amending its Form 10-K for the fiscal year 2006 and its Forms 10-Q for the first and second quarters of 2007.

In accordance with Section 404 of the Sarbanes-Oxley Act of 2002, our management assessed the effectiveness of our internal control over financial reporting. Based on both quantitative and qualitative factors, management has concluded that the findings detected during the investigation of the accounting for certain vendor program funds has resulted in the identification of a material weakness in internal controls over financial reporting. Management is evaluating and implementing changes in internal control over financial reporting relating to the timing of the recognition of vendor program funds in order to address the identified areas of the material weakness.

Evaluation of disclosure controls and procedures

Based upon the re-evaluation of the company's disclosure controls and procedures, as of the end of the period covered by this report, the company's principal executive officer and principal financial officer concluded that, as of such date, the company's disclosure controls and procedures were not effective at the reasonable assurance level, due to the fact that there was a material weakness in our internal control over financial reporting (which is a subset of disclosure controls and procedures) related to the timing of purchase commitments with vendors and the recognition of vendor program funds which resulted in the errors described in Note B to the consolidated financial statements. This material weakness resulted from deficiencies in the design of internal controls related to ensuring that complete and accurate documentation is provided to individuals responsible for the proper recognition of vendor program funds. The company's management recognizes that any controls and procedures, no matter how well designed and operated, can only provide reasonable assurance of achieving their objectives and management necessarily applies its judgment in evaluating the possible controls and procedures.

The company made no changes to its internal control over financial reporting for the quarter ended June 30, 2007. However, the material weakness discussed above was identified during 2007 and will result in future mitigation activities.

Item 1. Legal Proceedings

We are involved in litigation arising in the normal course of our business. While, from time to time, claims are asserted that make demands for large sums of money (including, from time to time, actions which are asserted to be maintainable as class action suits), we do not believe that any of these matters, either individually or in the aggregate, will materially affect our financial position or the results of our operations.

Office Depot has received a letter of informal inquiry from the United States Securities & Exchange Commission ("SEC"), looking into the company's contacts and communications with financial analysts during 2007. The company intends to cooperate fully with the SEC and does not anticipate commenting further on this matter while the inquiry is pending.

B) Q1/2006

Item 4. Controls and Procedures

(a) Disclosure Controls and Procedures. The company's management, with the participation of the company's Chief Executive Officer and Chief Financial Officer, has evaluated the effectiveness of the company's disclosure controls and procedures (as such term is defined in Rules 13a-15(e) and 15d-15(e) under the Securities Exchange Act of 1934, as amended (the "Exchange Act")) as of the end of the period covered by this report. Based on that evaluation, these officers have concluded that the corporation's disclosure controls and procedures are effective for the purpose of ensuring that material information required to be in this quarterly report is made known to them by others on a timely basis and that information required to be disclosed by the company in the reports that it files or submits under the Exchange Act is accumulated and communicated to the company's management, including its principal executive and principal financial officers, as appropriate to allow timely decisions regarding required disclosure.

(b) Changes in Internal Controls. The company is continuously seeking to improve the efficiency and effectiveness of its operations and of its internal controls. This results in refinements to processes throughout the company. However, there has been no change in the company's internal control over financial reporting that occurred during the company's most recent fiscal quarter that has materially affected, or is reasonably likely to materially affect, the company's internal control over financial reporting.

PART II. OTHER INFORMATION

Item 1. Legal Proceedings

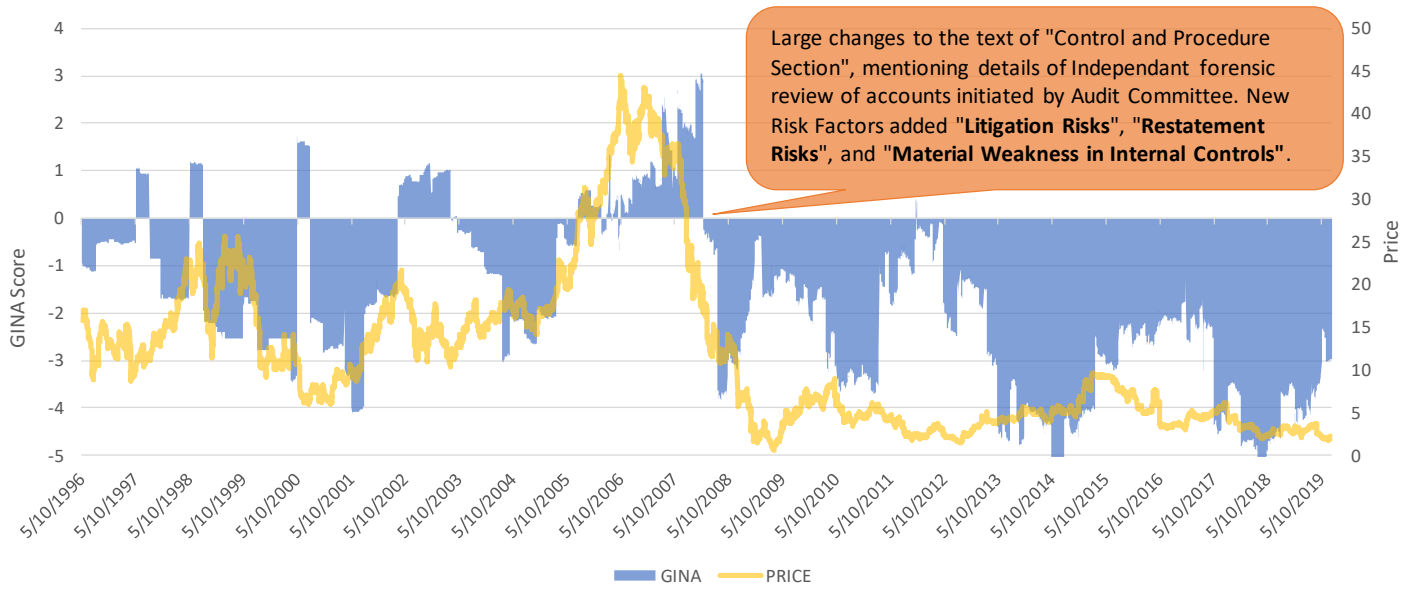
We are involved in litigation arising in the normal course of our business. While, from time to time, claims are asserted that make demands for large sums of money (including, from time to time, actions which are asserted to be maintainable as class action suits), we do not believe that any of these matters, either individually or in the aggregate, will materially affect our financial position or the results of our operations.

16

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

As shown in Figure 15, there was a sharp plunge exhibited by the GINA model (mostly triggered by the language similarity change) in 2007, which preceded the collapse of Office Depot's share price by three-to-six months. Based on the GINA model, the stock of Office Depot has remained in negative territory in the past decade, while the stock price has also stalled during the same period.

Figure 15 Changes in the "Risk Factors", "Control and Procedure", and "Legal Proceedings" Sections, Office Depot



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Wolfe Research Luo's QES

SENTIMENT ANALYSIS AND BEHAVIORAL FINANCE

In the world of NLP, sentiment or tone analysis is certainly one of the most widely used and exploited areas. In fact, the vast majority of data vendors in this space focus exclusively on sentiment analysis (e.g., RavenPack, Alexandria Technology, MarketPsych, Prattle Analytics, Accern).

Traditionally, sentiment analysis utilizes a pre-defined (either generic or context specific) dictionary with a pre-classified sentiment label for each word. The overall sentiment score of a document can be as simple as the percentage of positive (or negative) word frequency. Alternatively, more sophisticated weighting algorithms such as TF-IDF (Term Frequency-Inverse Document Frequency) are applied to each classified term.

Moreover, the strength of sentiment expressed in a text varies with the use of degree modifiers such as adjectives and adverbs. In recent years, a new generation of sentiment algorithms such as the VADER (Valence Aware Dictionary for sEntiment Reasoning) have gained tremendous popularity and shown great success. The VADER theory is explained in Hutto, et al [2014] and a practical implementation can be found in *Tone at the Top? Quantifying Management Presentation* (see Rohal, et al [2018]).

As detailed in *Beyond Fake News* (see Rohal, et al [2019]), traditionally, investors follow the "under-reaction to news" assumption. For example, when positive news is released to the market, not every manager is convinced. As a result, stock price appreciates, but not to the full extent justified by the news. Therefore, stock price continues to increase in the months subsequent to the news as investors play "catch up" to the earlier news. Our research indicates that the performance of the naïve buy-on-the-positive news (and sell-on-the-negative-news) strategy has retreated considerably in recent years. In our NICE (News with Insightful Categorical Events) model, we found that investors may either under-react (which is followed by conventional post-event-drift) or over-react (which leads to reversal) to news.

Using machine learning techniques, conditional on corporate events, the NICE model can effectively extract information from news and social media.

Sentiment analysis is further complicated by the nature of corporate regulatory filings. Unlike management presentations and conference calls, which are more bullish in nature, 10-K/10-Q filings are official documents filed with the regulators and heavily scrutinized by investors. Due to the fear of regulatory investigations and litigation, companies prepare such filings with great care. The language is generally conservative in nature and the documents are full of various disclosures. Other than the "MD&A" section, is the content is based more around facts and disclosures rather than opinions. Therefore, the information relevance of sentiment is very different than the sentiment in news, social media, or management presentations.

As a result, positive sentiment appears to be useful only in the "MD&A" section (see Figure 16A). On the other hand, more negative sentiment from the "Risk Factors" section is actually correlated to *higher* subsequent stock returns (see Figure 16B). More cautious language used in the "Risk Factors" section is possibly indicative of conservatism and rewarded by investors.

Figure 16 Positive and Negative Sentiment Factors (10-K Filings)

A) Positive Sentiment

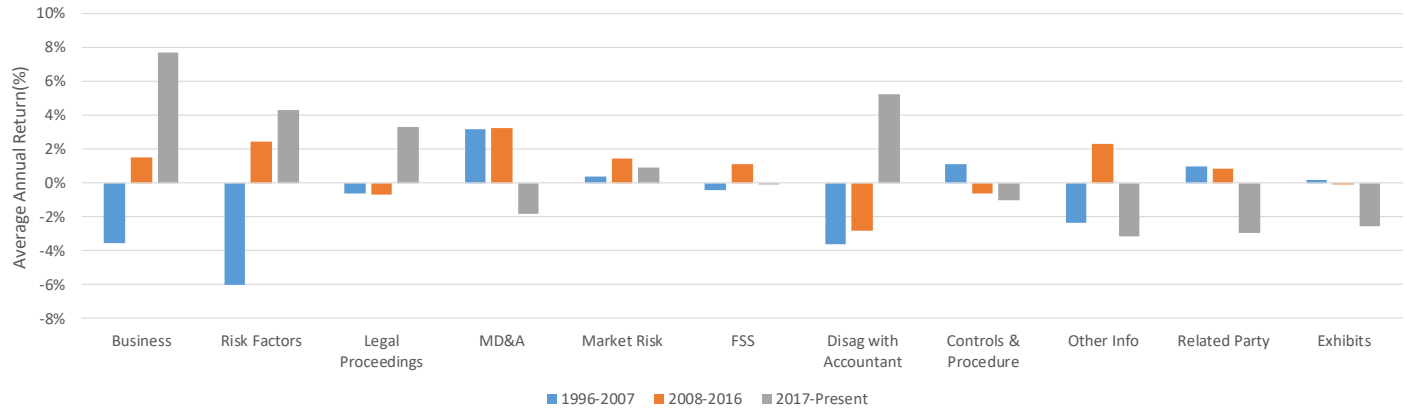
B) Negative Sentiment

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

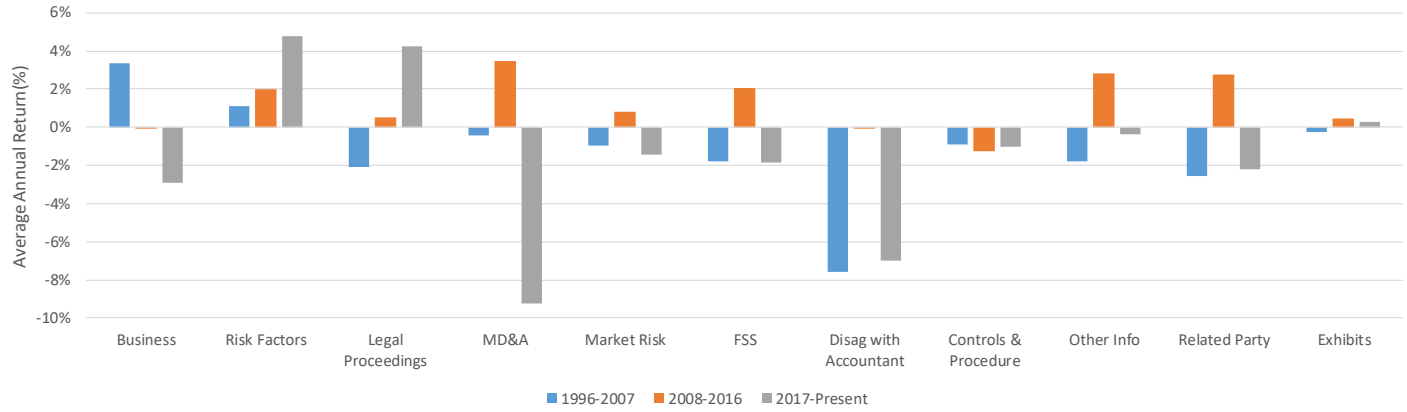
As shown in Figure 17, the performance of positive and negative signals is not very consistent – positive sentiment does not necessarily correspond to positive excess returns in the future, and vice versa for negative tone. Only in the case of the "MD&A" section where executives are more likely to offer their views of the future, the performance of positive and negative sentiment measures becomes more intuitive. In particular, companies with more negative tone in the "MD&A" section have materially underperformed in recent years.

Figure 17 Sentiment Measure Historical Performance (10-K Filings)

A) Positive Sentiment



B) Negative Sentiment



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo' s QES

Behavioral Finance

We can further analyze the structure of words based on formal grammar rules. In NLP terminology, this is called syntactic parsing. A syntactic parser reads an input sentence and describes its grammatical structure. The parser usually returns a graph of word-word relationships, aiming to extract the reasoning from the underlying text. One aspect of semantic analysis is part-of-speech (POS) tagging. It is the process of tagging words or tokens to their respective part of speech classes such as nouns, verbs and adjectives. These classes are known as lexical categories or parts of speech.

In language, adverbs modify verbs, while adjectives modify other nouns and pronouns. In presuppositions, even if you disagree with the adjective or adverb in the sentence (e.g., "This quarter's earnings were extremely strong."), you still accept what is modified (e.g., "This quarter's earnings were strong"). On the one hand, adjectives/adverbs generally enhance the baseline meaning. Using the VADER algorithm, for example, would increase the sentiment score. On the other hand, because adjectives and adverbs are less objective, the usage of these words in sentences measures the subjectivity of the document.

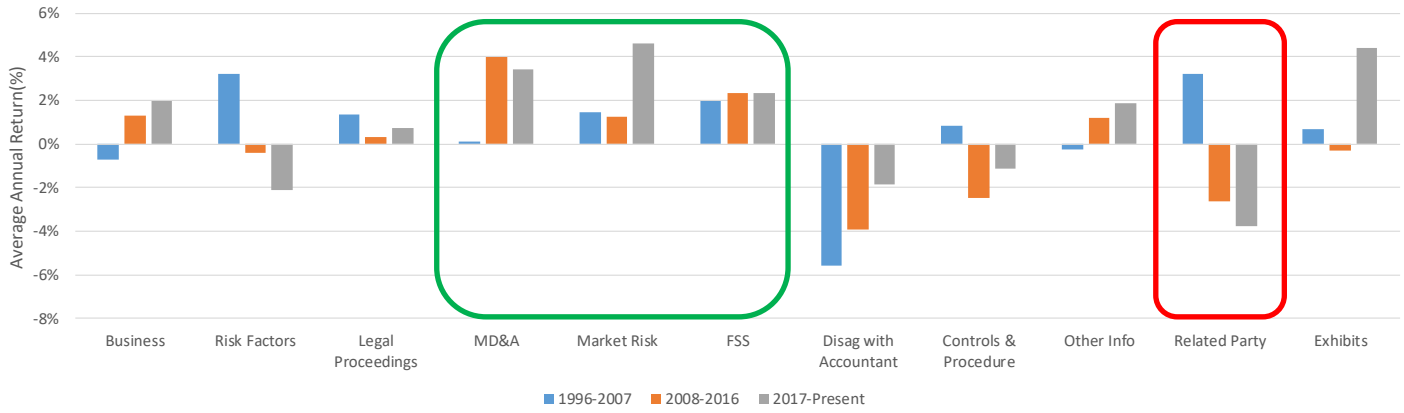
As explained in *Tone at the Top? Quantifying Management Presentation* (see Rohal, et al [2018]), another basic feature of management personalities is the handling of first person singular (I, me, my, mine) versus second person plural (we, us, our, ours) pronouns in the language. This is a subtle hue on the character of the executives as a team player. In management presentations, we found higher reference to self/first person is negative, while more use of team/second person pronouns is positively associated with future company performance.

The implications of adjectives/adverbs and first-person singular pronouns in corporate filings are very different from management conference calls. For example, as shown in Figure 18(A), a higher percentage of adjectives in the "MD&A", "Market Risk", and "FSS" sections is correlated to positive future stock returns, while the same factor in the "Related Party" section produces consistent negative returns.

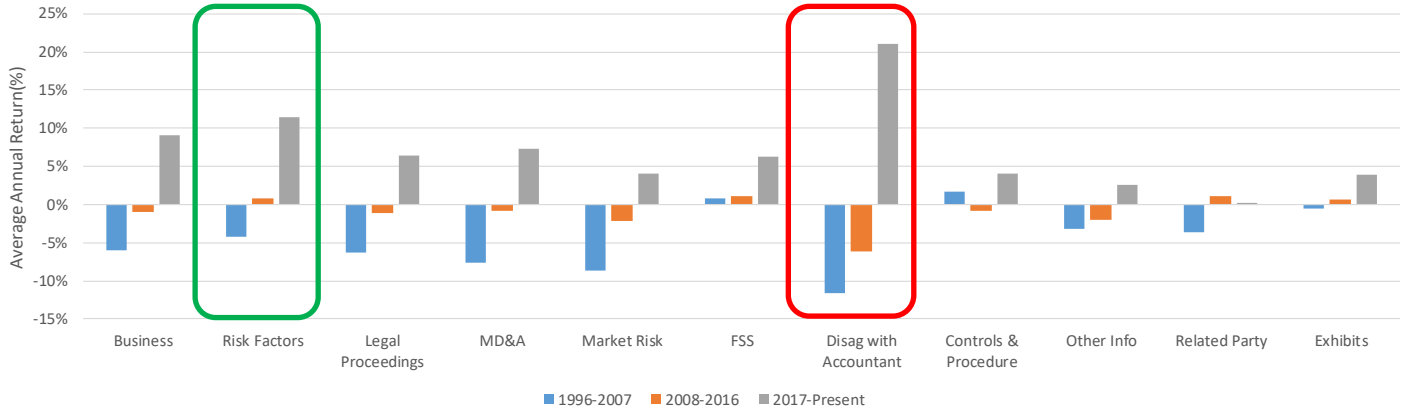
Similarly, how the market reacts to the usage of first-person singular pronouns is also drastically different from section to section. Apparently, investors seem to reward company management who is willing to take personal responsibilities in the "Risk Factors" section, but not so much in the "Disagreements with Accountants" (see Figure 18B).

Figure 18 Psychology-Based Factors (10-K Filings)

A) Use of Adjectives



B) Use of Pronouns



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

SentimentML – A Machine Learning Powered Sentiment Model

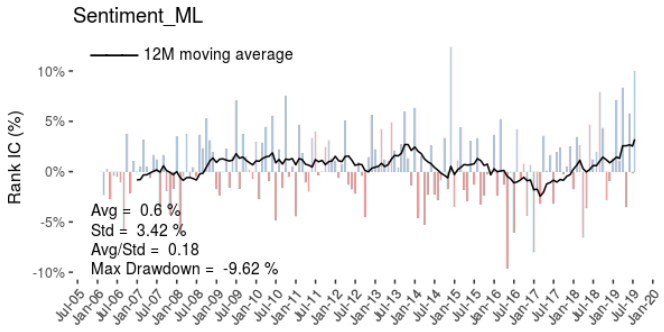
The language similarity factors measure the structural issues in company reporting. A large change in the structure in the disclosure sections (e.g., "Risk Factors") is indicative of future downside risk. The sentiment suite of factors, however, is rather different. It also depends on how the market interprets such sentiment and behavioral finance indicators. Therefore, the relationship between sentiment/behavioral finance factors and future stock returns is not necessarily linear and monotonic. Furthermore, as investors gradually learn how to read corporate filings and the market adapts, we are likely to see a time varying relationship.

As argued in *Beyond Fake News* (see Rohal, et al [2019]), machine learning techniques are well designed in areas where the patterns are highly nonlinear and time varying. Similar to the NICE model, we apply a combination of Elastic Net and xgBoost to capture both linear and non-linear relationships behind the sentiment and behavioral finance signals. The SentimentML model is trained on a 10-year rolling window to predict one-quarter ahead stock returns.

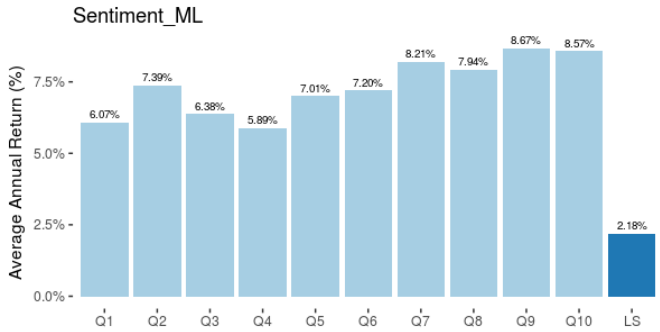
As shown in Figure 19(A), the SentimentML model delivers a reasonably consistent performance in the past decade, with a long forecasting horizon beyond a year (see Figure 19B). A simple long/short quintile portfolio based on the model generates an annual Sharpe ratio of 0.5x (see Figure 19C).

Figure 19 Performance of the Composite Sentiment Model (10-K and 10-Q Filings)

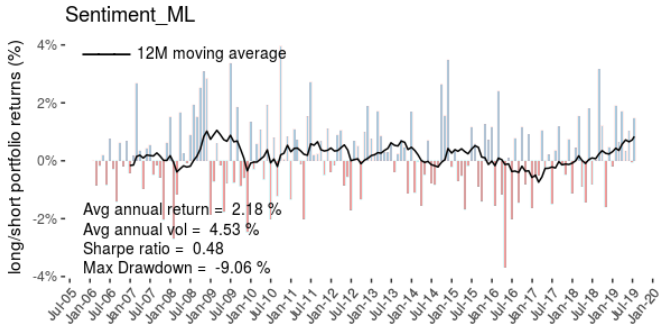
A) Rank IC



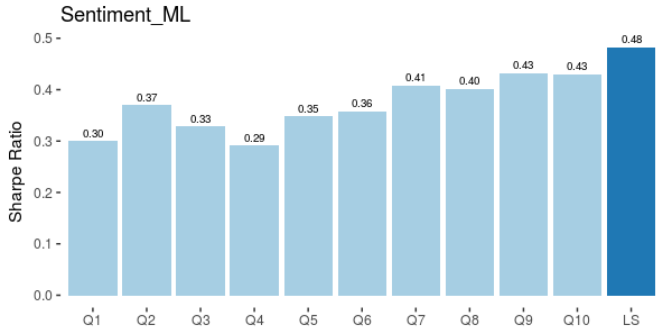
B) Annualized Portfolio Return



C) Long/Short Decile Portfolio Return



D) Sharpe Ratio



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

THE NEXT FRONTIER – CNN (CONVOLUTIONAL NEURAL NETWORKS)

In March 2016, AlphaGo – a deep learning powered computer program – beat Lee Sedol (a Ninth Dan professional Go player) in the board game Go. Then in 2017, AlphaGo further beat Ke Jie – the world's No.1 ranked player at the time. Although computer algorithms have dominated chess since 1996, it was long been considered as impossible to beat human players in the Go game. Since the spectacular success of the AlphaGo program, deep learning or deep neural networks have gained tremendous popularity in the AI and machine learning world.

In this section, we explore one particular technique in the deep learning field – CNNs (Convolutional Neural Networks) in processing corporate regulatory filings.

What is an Artificial Neural Network?

Artificial Neural Networks (ANNs) are computing systems that are inspired by biological neural networks that constitute human brains. An ANN is composed of many highly interconnected neurons working in unison to solve a specific problem such as image recognition or text classification. An ANN can be used to extract patterns and detect trends that are too complex to process by other ML techniques. There are a variety of ANNs that operate in their specific ways to achieve different outcomes.

Due to the huge demand on computing power, traditional ANNs typically consist of an input layer, a hidden layer, and an output layer. In recent years, with the advance in both numerical algorithm designs and brute force computing power, CNNs and RNNs (Recurrent Neural Networks) type of deep neural networks with multiple hidden layers have gained wide acceptance.

Data Cleaning and Transformation

Before fancy NLP and ML algorithms can be deployed for text mining, data cleaning and transformation is a critical step. The text corpus needs to be cleaned and organized before training a CNN model. Although these efforts vary with use cases, the basic steps are similar:

- Converting text to lower case, removing extra white spaces and line breaks
- Removing numbers, punctuations marks and special characters; if needed, special tokens can be assigned to signify numbers or special characters
- Removing stop-words (e.g., 'the', 'is', 'are', 'a', 'an'), other common words (e.g., proper nouns such as January, London), which are irrelevant to our research
- Padding each document to the maximum document length, by adding special </PAD> tokens to all documents, so they have the same size[5]
- Building a vocabulary index and mapping each word/token to an integer between zero and the vocabulary size, so each document becomes a vector of integers
- Transforming each word in the document into a vector of numbers called "word embeddings" or "word vectors", which are then used in the CNN model
- Stemming or lemmatization of words to their roots are explored

[5] Padding documents to the same length allows us to efficiently batch our data, because each example in a batch must be of the same length.

- Assigning a label (positive/negative) to each document for training[6]

CNN INTRODUCTION

CNNs are regularized versions of multilayer deep ANNs. In fully connected multilayer ANNs, each neuron in one layer is connected to all neurons in the next layer. The full connections make these classes of ANNs prone to overfitting. CNNs take an interesting approach to regularization, by taking advantage of the hierarchical pattern in data, inspired by the animal visual cortex. CNNs require minimal pre-processing compared to other classification algorithms. CNNs learn the filters that are often hand-engineered in traditional algorithms. As a result of the regularization and absence of pre-processing, CNNs are robust to overfitting[7]. Traditionally, CNNs were mostly used in computer vision. However, given their successes in image recognition, CNNs have found their way into NLP[8].

Model Training and Prediction

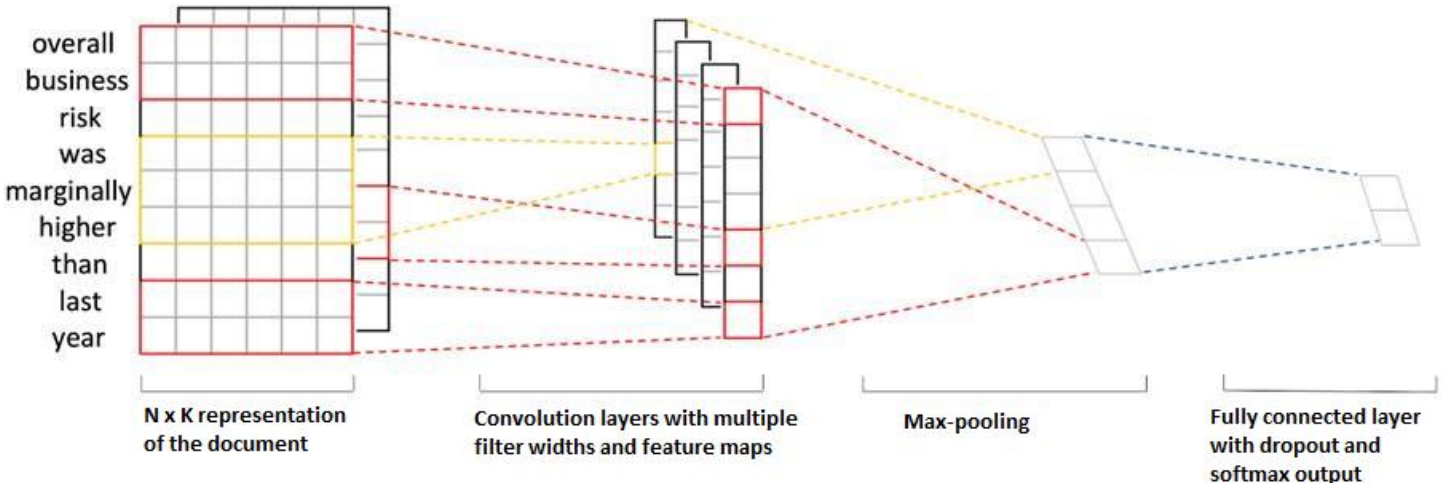
Figure 20 shows a representative diagram of our implementation of CNNs in textual analysis of corporate filings. Our CNN also consists of multiple layers of perceptrons. The first layer conducts word embedding into low-dimensional vectors (128 dimensions). It is important to note that the word vectors in our CNN are not pre-trained. Instead, the model learns embeddings from scratch, which ensures that the embeddings are in context and related words in 10-K/10-Q filings have similar embeddings.

The next multilayers perform convolutions over the embedded word vectors using multiple filter sizes, e.g., sliding over two, three, or five words at a time. By varying the size of the kernels and concatenating their outputs, we can detect patterns of multiples sizes (i.e., two, three, or five adjacent words). Identified patterns could be expressions such as “margin compression”, “cost pressure”, “favorable pricing”. CNN models can identify these informative patterns in sentences regardless of their positions.

Next, we max-pool the result of the convolutional layers into a long feature vector, add dropout regularization, and classify the result using a softmax layer. Dropout is perhaps the most popular method to regularize CNNs. The idea behind dropout is simple. A dropout layer stochastically “disables” a fraction of its neurons, which prevents neurons from co-adapting and forces them to learn useful features independently.

[6] In our case, labels are based on 12-month forward stock returns (normalized for beta, size and sector).
[7] In a similar philosophy, we have found our MBBT (Multi-Branch Boosted Tree) algorithm to be robust to overfitting in stock selection (see *Man versus Machine – MALTA*, Wang, et al [2018] for details).
[8] Interestingly, the predecessor of our MBBT algorithm – the AdaBoost technique – was also proven to be highly effective in image processing first. We adopted the AdaBoost algorithm in stock selection in 2012.

Figure 20 Representation of a Convolutional Neural Network (CNN) Model



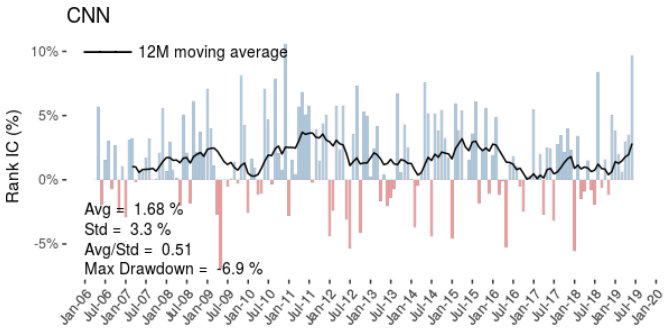
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

One particular drawback of training a CNN model on raw text documents is its computational intensity – even with our large-scale GPU/CPU parallel computing platform. Since the number of features (i.e., factors) equals the number of unique words in our dictionary, our model constantly seeks more data. We focus our attention on a few of the most informative sections in the 10-K filings, e.g., "Business", "Risk Factors" and "MD&A".

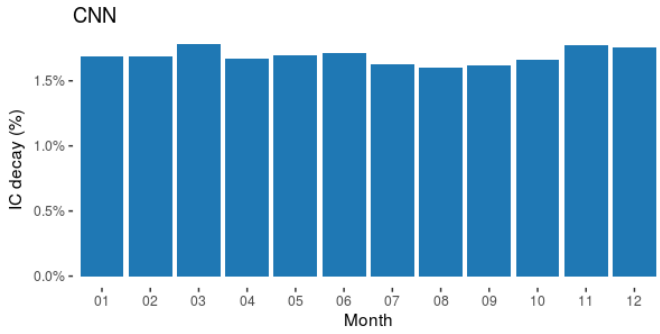
The CNN model training on the "MD&A" section delivers a similar Sharpe ratio to our SentimentML, but beats language similarity-based factors (see Figure 21). More importantly, as more and more companies adopt HTML tagging in their EDGAR filings, our QES Scraper is able to extract more and more companies with complete mappings. As data quality improves over time, the performance of our CNN model also gets better (see Figure 21A). Lastly, similar to most other NLP signals introduced in this paper, the CNN model also has a long investment horizon and slow signal decay (see Figure 21B)

Figure 21 CNN Model Trained Using the "MD&A" Section (10-K Filings)

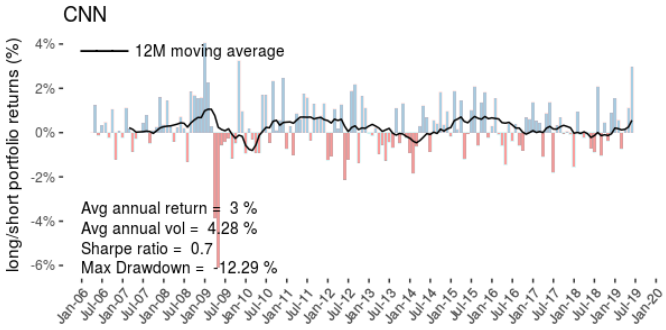
A) Rank IC



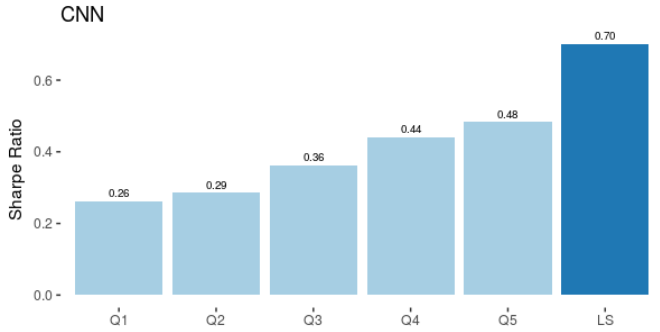
B) IC Decay



C) Long/Short Quintile Portfolio Return



D) Sharpe Ratio

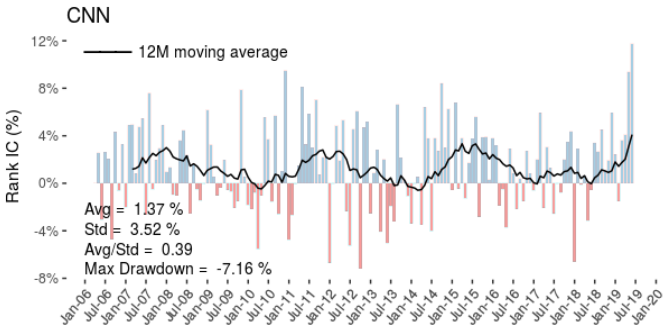


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

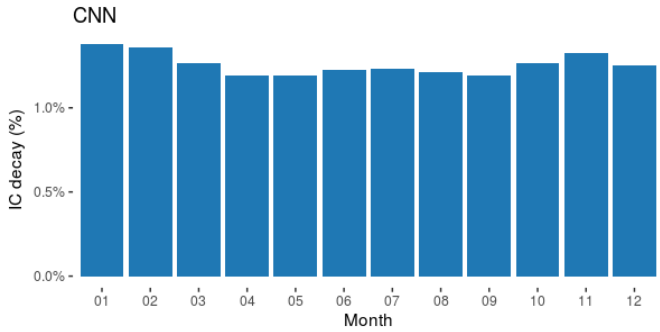
As a robustness test, Figure 22 shows the performance of the CNN model trained using the "Business" section. The results are largely the same.

Figure 22 CNN Model Trained Using the "Business" Section (10-K Filings)

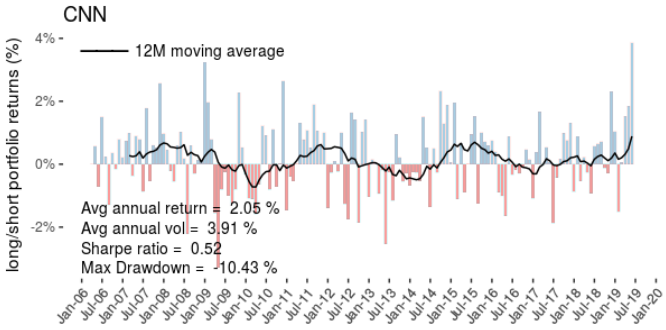
A) Rank IC



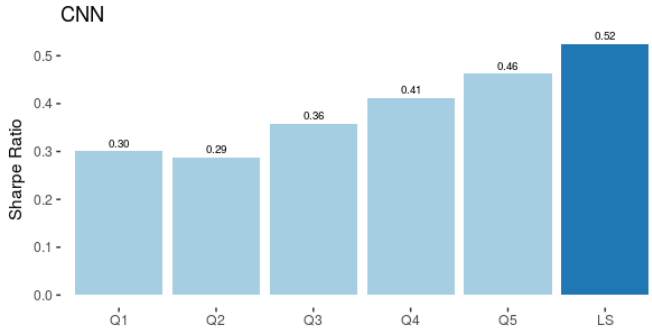
B) IC Decay



C) Long/Short Quintile Portfolio Return



D) Sharpe Ratio



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

INTRODUCING THE GINA (US) MODEL

Finally, we combine the signals and models discussed in this paper together and introduce the fifth generation of our NLP/ML model – the GINA (Global Intelligence NLP Alpha). The GINA is based on our previous NLP model for corporate filings – the SPEC – using our latest technology infrastructure and NLP/ML algorithms, and supersedes the SPEC model.

Essentially, the GINA model has four underlying components:

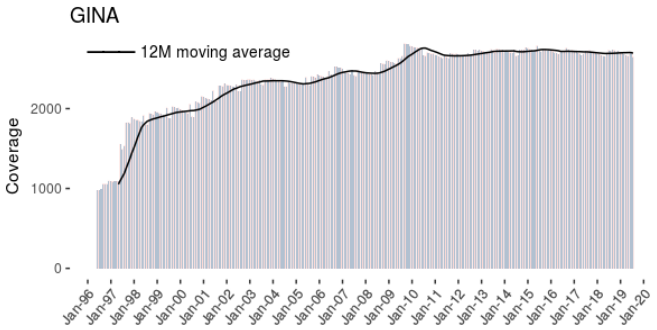
- Two language similarity composites – one based on the various sections in the 10-K and another one learned from the 10-Q filings
- SentimentML – the machine learning based sentiment model, using the Elastic Net and xgBoost algorithms on both sentiment/tone factors and behavioral finance motivated signals
- CNN model – a model leveraging the CNN deep learning algorithm, trained on the original words in the filings without pre-processing

The GINA model covers almost 90% of the Russell 3000 universe (see Figure 23A). Measuring by rank IC, the GINA model has shown consistent predictive power of future stock returns in the past 20 years (see Figure 23B), with long investment horizon/slow decay (see Figure 23C). A long/short decile portfolio based on the model generates an average annual return of 8% (see Figure 23C), with a Sharpe

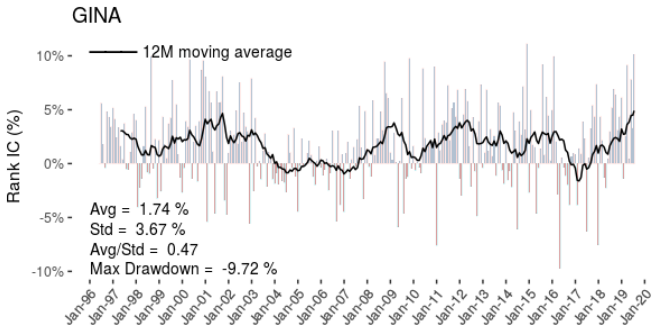
ratio of 1.0x (Figure 23E). The model payoff pattern is monotonic across the 10 decile portfolios (see Figure 23E). Lastly, the turnover of the GINA model is fairly modest, much lower than most quantitative models and sentiment-based signals. The spike in turnover around 2006 is due to SEC's overhaul in the EDGAR filing template.

Figure 23 GINA(US) Model Performance

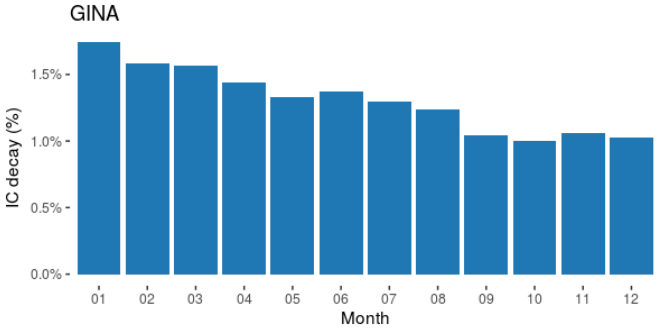
A) Coverage



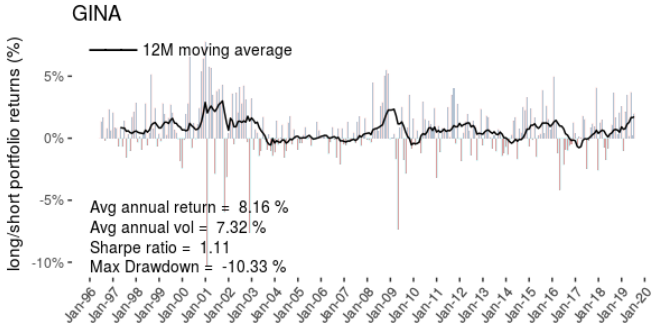
B) Rank IC



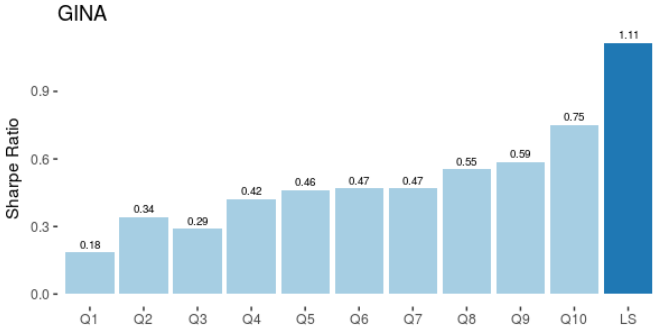
C) IC Decay



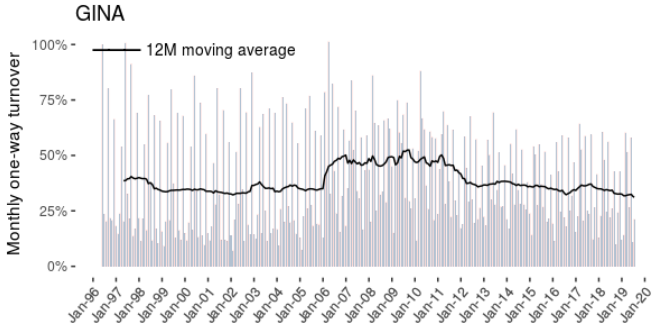
D) Long/Short Decile Portfolio Return



E) Sharpe Ratio



F) Turnover



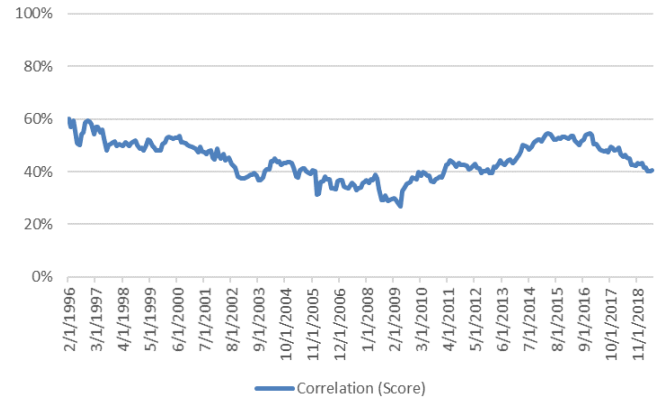
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

Performance Comparison – GINA versus SPEC

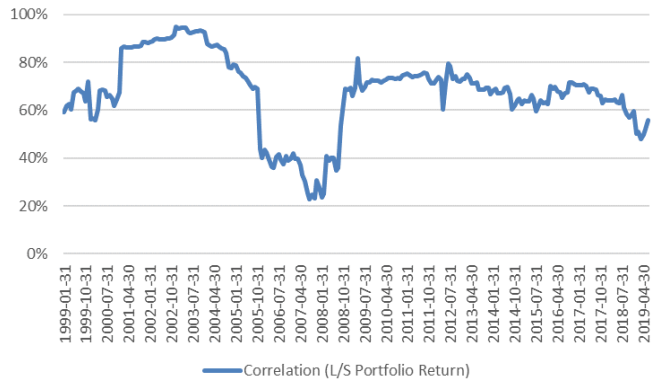
Although the GINA model can be perceived as an upgrade of the existing SPEC model, the underlying NLP/ML algorithms are meaningfully different. Therefore, investors could use both models at the same time. As shown in Figure 24(A), the model score correlation, i.e., the correlation between the GINA model ranking and the SPEC model output, is only around 40% on average. Similarly, based on performance of the long/short decile portfolio formed on the two models, the correlation is also fairly modest, roughly 70% on average (see Figure 24B).

Figure 24 Model Correlation – GINA versus SPEC

A) Model Score Correlation



B) Performance Correlation



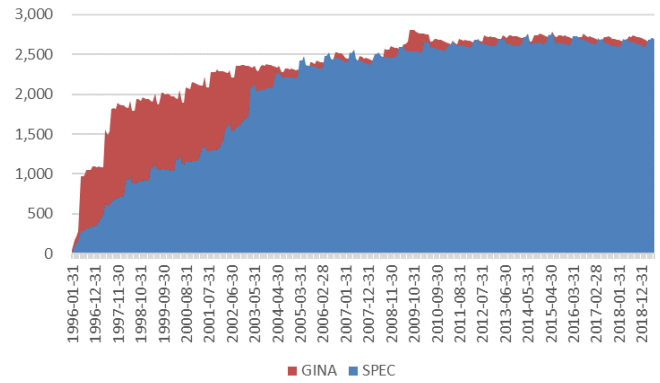
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo' s QES

In terms of model coverage, the GINA is almost 15% better than the SPEC, especially during 1990s and early 2000s (see Figure 25A). The GINA model delivers a Sharpe ratio (based on a long/short quintile portfolio) that is 32% higher than the SPEC (see Figure 25B), with a much lower downside risk – the maximum drawdown is 41% lower (see Figure 25C). The turnovers (signal autocorrelation) of the two models are comparable (see Figure 25D).

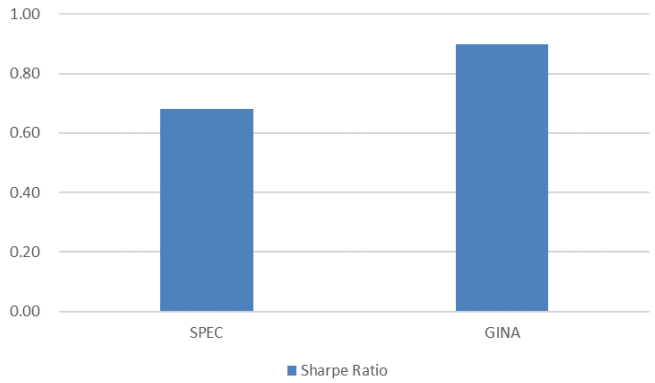
More importantly, unlike most traditional stock-selection factors whose performance has suffered tremendously in recent years, the performance of both SPEC and GINA has in fact improved since 2017 (see Figure 25E). The Sharpe ratio of the GINA model since 2017 is more than 60% higher than its long-term average. Similarly, the downside risk of both models has also moderated in recent years (see Figure 25F). The worst monthly return of the GINA since 2017 was merely -3.1%.

Figure 25 Performance Comparison – GINA versus SPEC

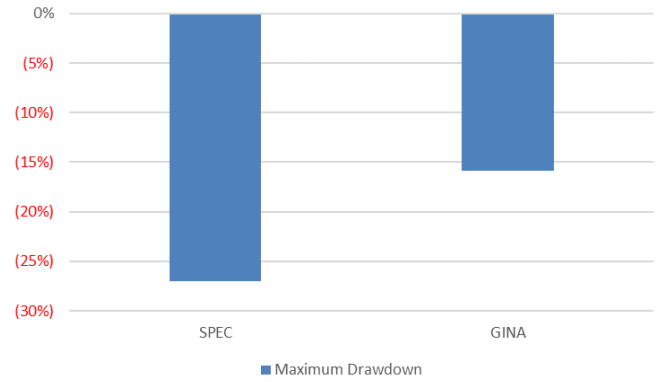
A) Coverage



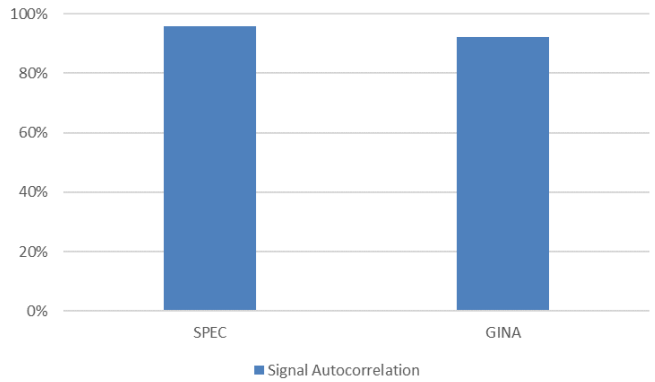
B) Sharpe Ratio (Long/Short Portfolio), 1996-2019



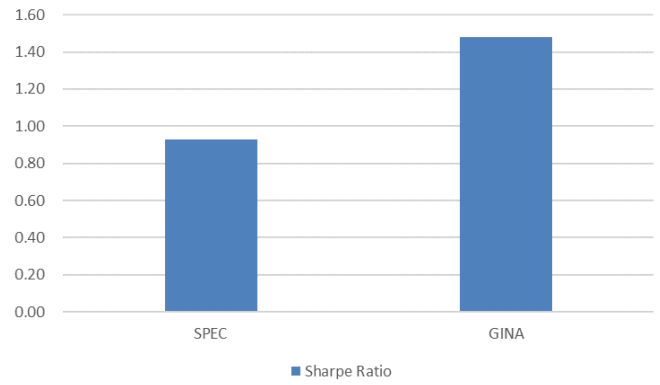
C) Maximum Drawdown, 1996-2019



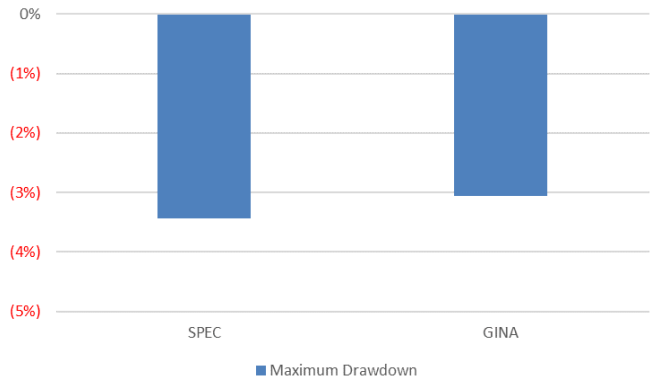
D) Signal Autocorrelation (Monthly Average)



E) Sharpe Ratio, 2017-2019



F) Maximum Drawdown, 2017-2019



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo' s QES

TEXT MINING INTERNATIONAL ANNUAL REPORTS AND INTERIM FILINGS

So far, we have discussed our fifth generation of NLP model – the GINA and how the model analyzes 10-K/10-Q filings from the EDGAR database. In this section, we extend our GINA model to international companies.

There are several major challenges when we extend the model from the US to the international market. First, in the US, public companies have to file their documents according to the prescribed format mandated by the SEC. The EDGAR database contains over 20 years of filing data. Outside of the US, however, there is no standard filing requirements. Although data vendors such as Compustat, S&P Capital IQ, and Worldscope collect and standardize financial statement data, the textual information such as annual reports and interim statements are rarely being systematically analyzed. For example, in Canada, public company filings are available from the SEDAR (System for Electronic Document Analysis and Retrieval) platform, which provides a semi-standardized system of company filings. Annual reports, interim filings, and key sections such as MD&A can be downloaded as PDF or text files. However, there is no standardized HTML tagging. In most other countries, standard corporate filing databases do not even exist.

Second, although financial statement data is often translated in English and reconciled across different accounting standards, textual documents and disclosures can be written in many languages. English version annual and interim reports are either not available, or available with significant delays. The vast majority of existing NLP algorithms are designed for English and English only. Extending English-language NLP to other languages is a daunting task.

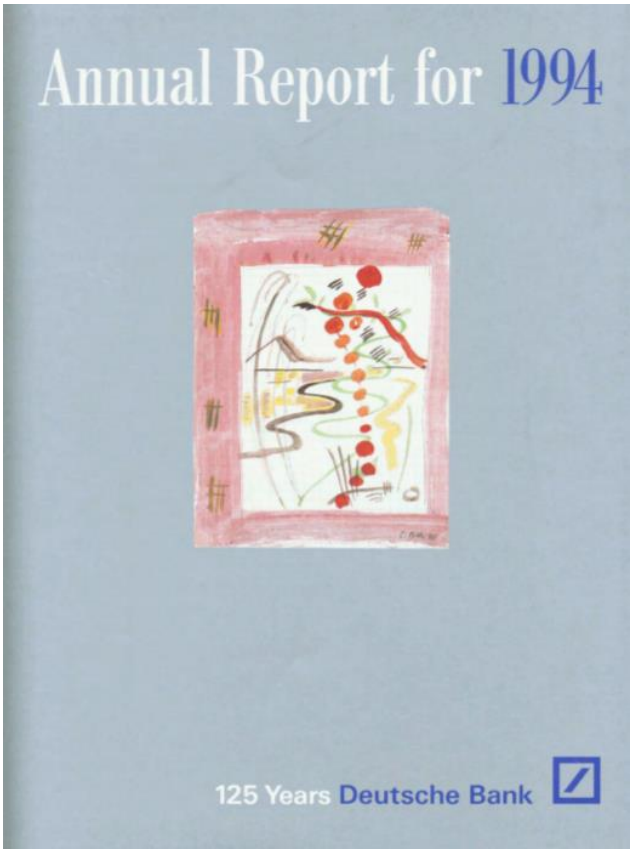
Furthermore, culture and corporate governance structure can also be vastly different in different countries. For example, compared to the US, Japanese companies tend to emphasize more on team decision, conservative corporate style, more on the quality of products and services and less on financial performance. In addition, cross-shareholding is commonplace in Japan, but rare among US firms. Clearly, even the same words (after translation) may convey very different meanings.

The Mergent Corporate Filing Database

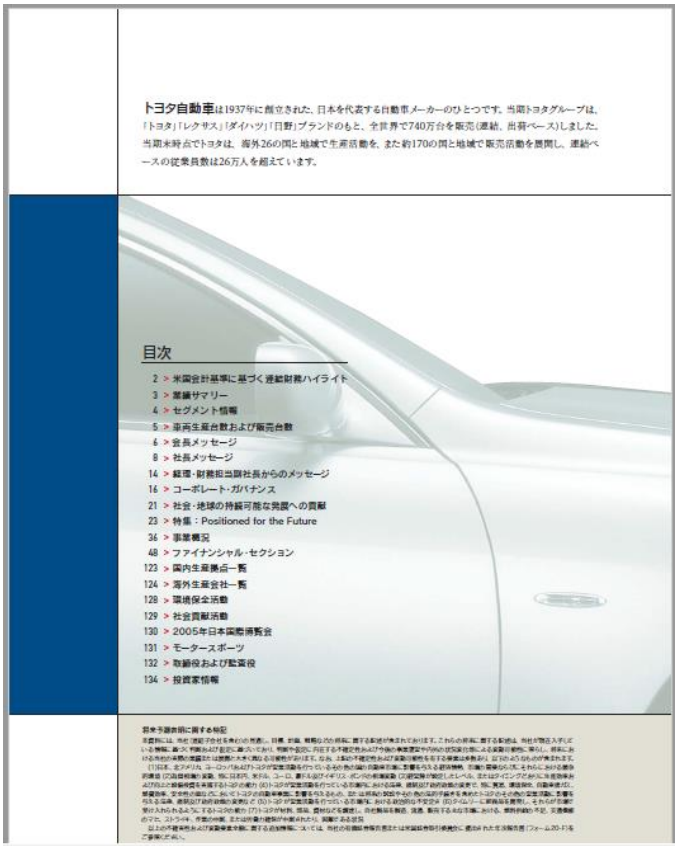
For non-US corporate filings, we source our data from Mergent – a FTSE Russell company. Mergent has collected an encyclopedia of annual reports and interim filings for most public companies globally, going back almost 50 years. The files are in PDF format. In the earlier years, documents are mostly scanned images, e.g., Deutsche Bank in 1996 (see Figure 26A), while in recent years, the filings are mostly text-based PDF, e.g., Toyota in 2005 (see Figure 26B). While many large multinational companies produce annual reports in English, local language version is more widely available on a more timely fashion (see Figure 26B).

Figure 26 Examples of Annual Reports

A) Deutsche Bank (1994) Cover Page



B) Toyota (2005) Table of Contents

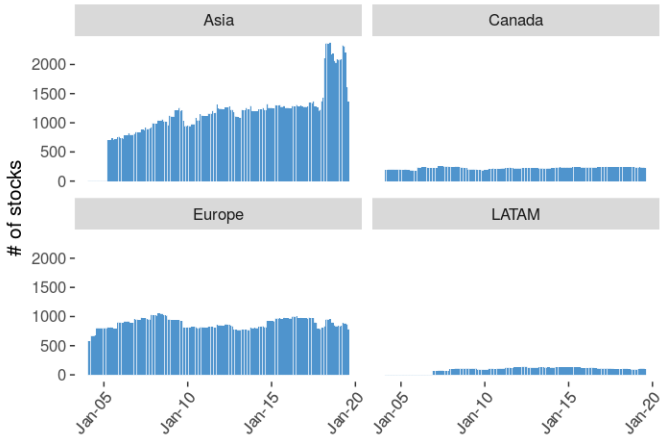


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

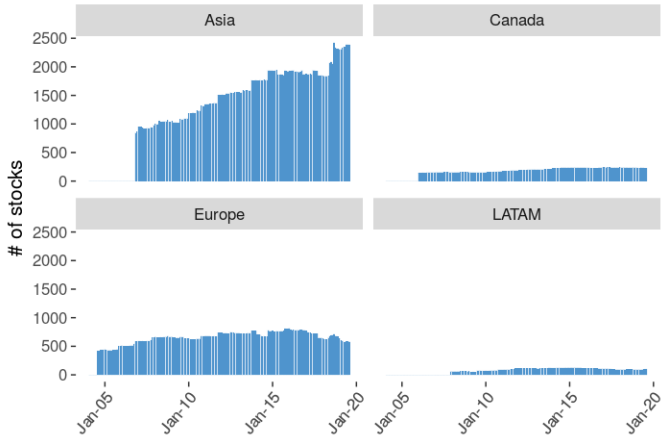
The Mergent database is divided into four regions outside of the US: Canada, LATAM (Latin America), Asia (including China and Japan), and Europe. We focus our attention on the Mergent PDF filings with embedded text. In other words, the scanned image version of PDF documents is not used in this research. For the PDF filings where we could extract the underlying text directly, we get meaningful coverage for both the annual and interim filings in most regions since 2005 (see Figure 27). Coverage in Latin America is rather limited. Moreover, interim filings seem to be only widely available in Asia.

Figure 27 International Coverage

A) Annual Reports



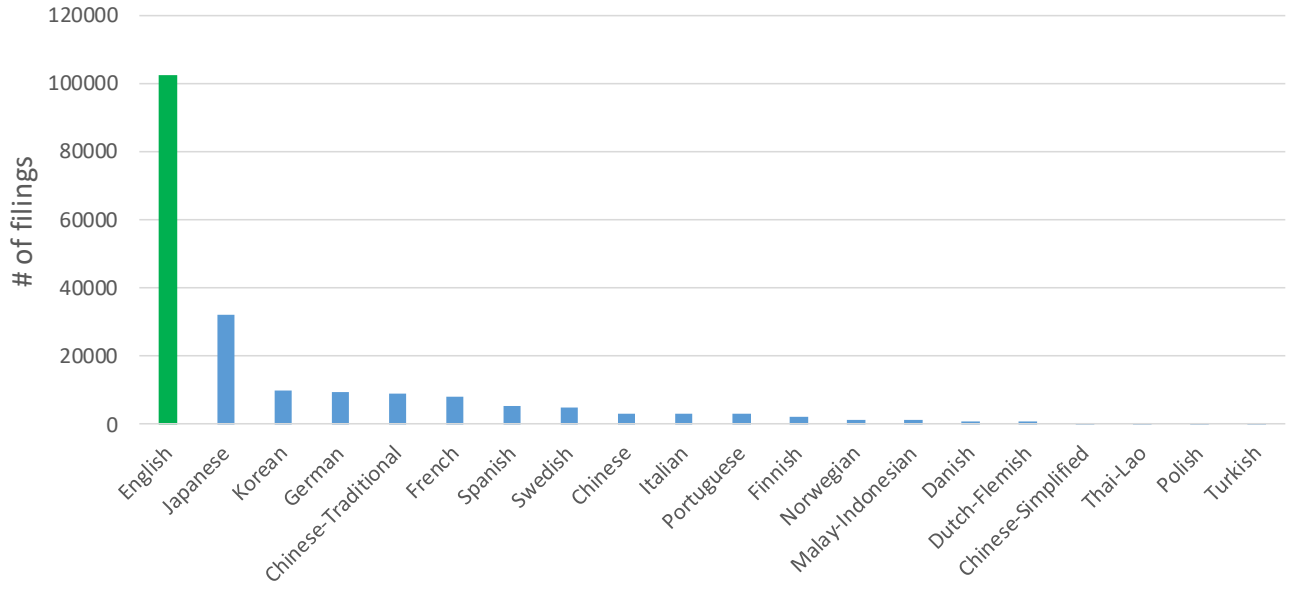
B) Interim Filings



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

Based on the Mergent database, most international companies release annual and interim reports in English (see Figure 28). Most European companies file financial documents in English, although German, French and Spanish are also widely used. In Asia, most Japanese and Korean filings are in their native language, while firms in Hong Kong, Singapore, and other Asian countries use predominantly English.

Figure 28 # of Filings, by Language



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

TOKENIZING NON-ENGLISH LANGUAGES

Before we can apply NLP/ML on the underlying text, one of the crucial steps is to tokenize words embedded in textual documents. For English and other European languages, tokenization is a simple process, given the word boundaries are made of simple delimiters like comma and spaces. For some Asian languages, in particular, Japanese, Chinese, and Korean, there are no such clear word boundaries, which requires language specific tokenizers. Tokenizing these three most prominent languages can be achieved with language specific libraries. For example, Jieba, tinysegmenter, and koNLPy are the three Python libraries that can be used to tokenize Chinese, Japanese, and Korean languages, respectively.

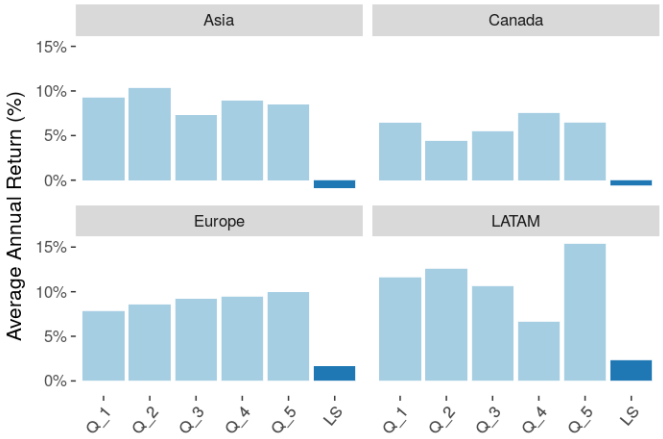
LANGUAGE SIMILARITY SIGNAL

Since the international corporate filings are highly unstructured without a pre-defined set of sections, our similarity based factors can only be computed on the entire document. Once we have the tokenized text, we apply the same techniques, e.g., Cosine and Jaccardian similarities used in GINA(US) for international companies.

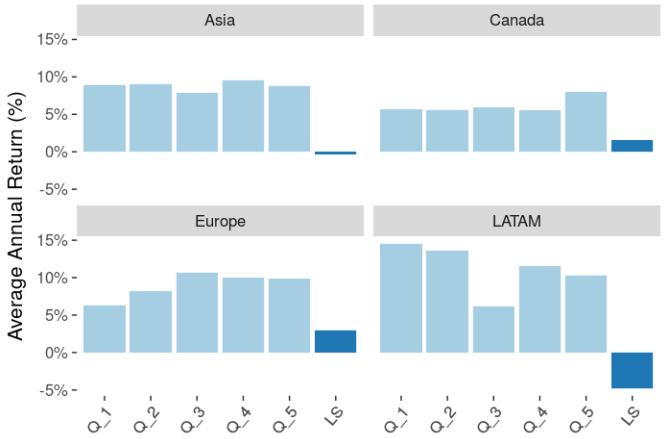
Using annual reports, the language similarity measures are effective in predicting future stock returns in Europe and LATAM, but less so in Asia and Canada (see Figure 29). With the more timely interim filings, the performance of our signals improves considerably (see Figure 29), which is consistent with what we observed in the US. The less impressive results in Canada and LATAM is partially due to the small sample and lack of coverage for interim filings.

Figure 29 Similarity Measure Performance

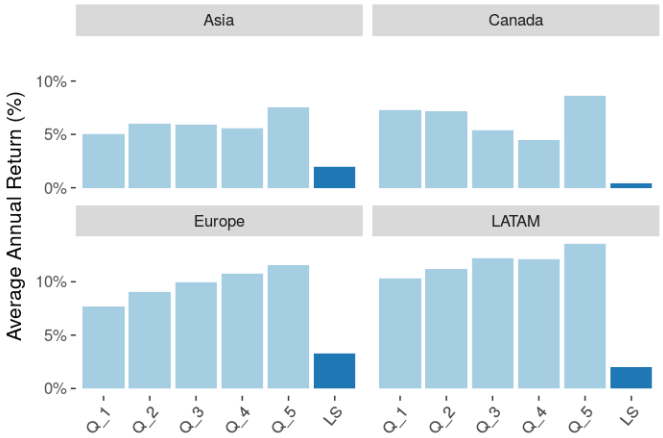
A) Cosine Similarity, Annual Reports



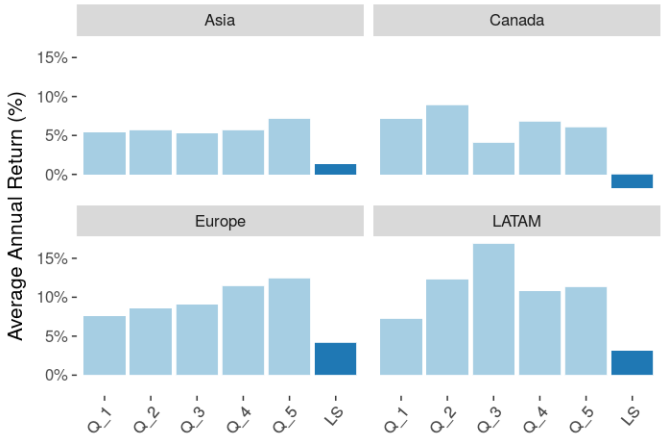
B) Jaccardian Similarity, Annual Reports



C) Cosine Similarity, Interim Filings



D) Jaccardian Similarity, Interim Filings

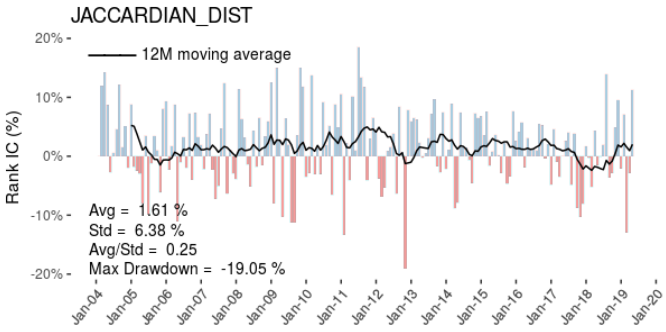


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo' s QES

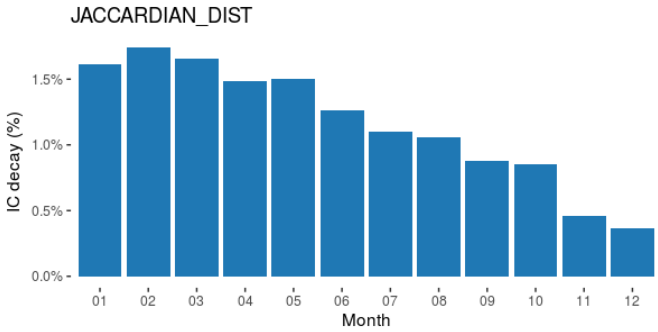
Figure 30 shows the historical rank IC of the Jaccardian similarity factor (interim filings) in Europe and Asia. The average IC of 1.0%-1.5% is roughly in-line with the US. The decay profile in Europe and Asia is slightly faster than in the US, but the similarity factor still shows reasonably long investment horizon (see Figure 30B and D).

Figure 30 Performance of YoY Jaccardian Similarity Measure (Interim Filings)

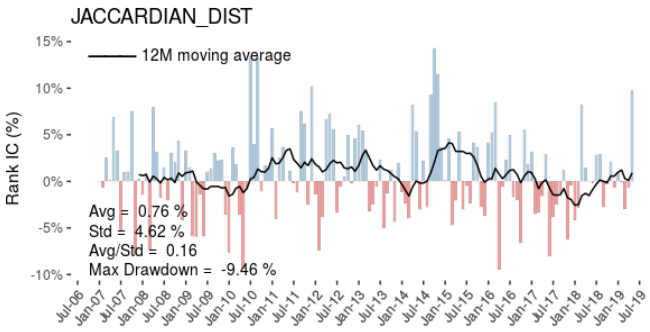
A) Rank IC (Europe)



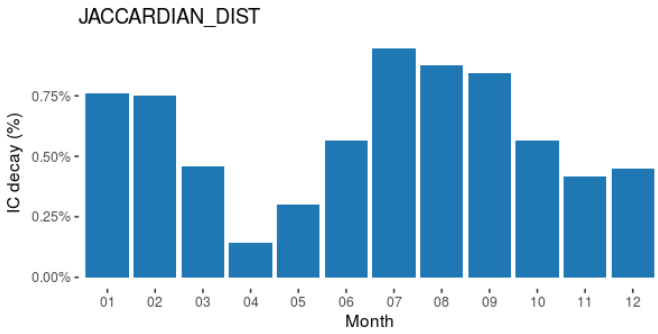
B) IC Decay (Europe)



C) Rank IC (Asia)



D) IC Decay (Asia)

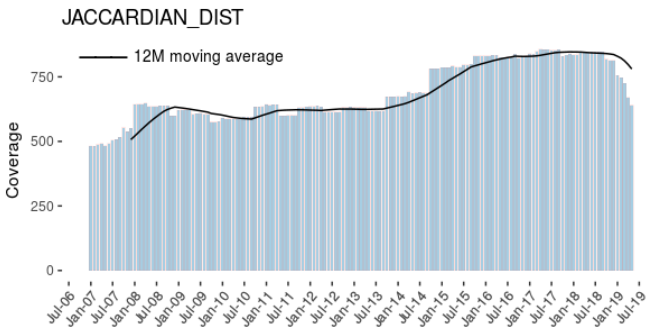


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

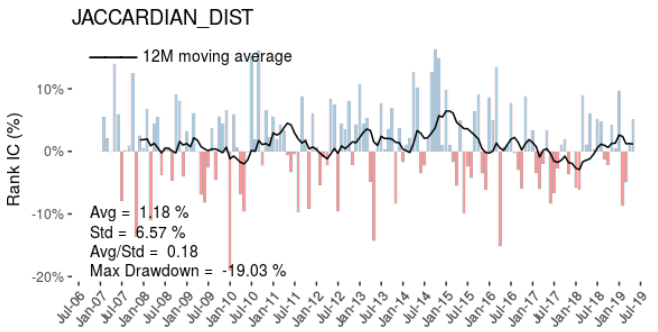
Given the unique nature of the Japanese market and the large number of public companies in Japan, we can also separately backtest the Jaccardian similarity factor in the Japanese markets. With more than 700 companies in Japan (see Figure 31A), the performance of the similarity measure is slightly better in Japan than the combined Asian market.

Figure 31 Performance of YoY Jaccardian Similarity Measure in Japan (Annual Reports)

A) Coverage



B) Rank IC



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

SENTIMENT ANALYSIS AND BEHAVIORAL FINANCE

For the purpose of sentiment/tone analysis, to process the large number of non-English filings, we have created a multilingual sentiment dictionary of 21 different languages. A sample of the dictionary is shown in Figure 32. Words in the annual and interim filings are then assigned sentiment as either positive or negative based on our multilingual dictionary.

Figure 32 A Set of Sample Words from a Multilingual Dictionary

English	Swedish	Chinese	Korean	French	German	Japanese	Russian	Thai.Lao	Portuguese	Spanish
achieve	uppnå	实现	이루다	atteindre	leisten	達成する	достигать	บรรลุ	conquistar	lograr
bad	dålig	坏	나쁜	mal	Schlecht	悪い	Плохо	ไม่ดี	mau	malo
challenging	utmanande	具有挑战性的	도전적인	difficile	herausfordernd	挑戦	испытывающий	การท้าทาย	desafiador	desafiante
denial	avslag	否认	부정	le déni	Verweigerung	拒否	отказ	การปฏิเสธ	negação	negación
efficacy	effektivitet	功效	효능	efficacité	Wirksamkeit	効能	эффективность	ประสิทธิภาพ	eficácia	eficacia
excellent	excellent	优秀	우수한	excellent	Ausgezeichnet	優れた	отлично	ยอดเยี่ยม	excelente	excelente
excited	upphetsad	兴奋	흥분한	excité	aufgeregt	興奮した	в восторге	ตื่นเต้น	animado	emocionado
good	Bra	好	좋은	bien	gut	良い	хорошо	ดี	Boa	bueno
great	bra	大	큰	génial	groß	すばらしいです	Великий	ยิ่งใหญ่	ótimo	estupendo
insecure	osäker	不安全	불안정한	peu sûr	unsicher	安全でない	небезопасный	ไม่ปลอดภัย	inseguro	inseguro
insolvency	insolvens	破产	지불 불능	insolvabilité	Insolvenz	倒産	неплатежеспособность	การล้มละลาย	insolvência	insolvencia
jury	jury	陪审团	배심	jury	Jury	陪審	жюри	คณะลูกขุน	júri	jurado
opportunity	möjlighet	机会	기회	opportunité	Gelegenheit	機会	возможность	โอกาส	oportunidade	oportunidad
positive	positiv	正	양	positif	positiv	ポジティブ	положительный	บวก	positivo	positivo
prosecution	åtal	检察官	기소	poursuite	Strafverfolgung	起訴	судебное преследование	การฟ้องร้อง	acusação	enjuiciamiento
risk	risk	风险	위험	risque	Risiko	リスク	риск	อันตราย	risco	riesgo

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

In linguistics, there are around 7,200 languages in the world, while about 800 of them are considered as "main" languages. Even the 21 languages in our example belong to many branches and groups (see Figure 33). While the Indo-European family of languages share some commonalities, the Asian languages such as Japanese, Korean, Sino-Tibetan (Chinese and Thai-Lao) require very different dictionaries and tokenizers to analyze.

Figure 33 World Language Classification System

Family	Group	Language	#
Indo-European			
	Portuguese	Portuguese	1
	Spanish	Spanish	2
	French	French	3
	Italian	Italian	4
	Germanic	Germanic	
		English	5
		Dutch-Flemish	6
		German	7
		Danish	8
		Swedish	9
		Norwegian	10
	Slavic		
		Polish	11
		Russian	12
Uralic			
	Finno-Ugric		
		Finnish	13
Altaic			
	Turkic		
		Turkish	14
Japanese			15
Korean			16
Sino-Tibetan			
	Chinese		
		Chinese	17
		Chinese (Mandarin-Simplified)	18
		Chinese (Cantonese-Traditional)	19
	Tai-Kadai		
		Thai-Lao	20
Austronesian			
	Western Malygo-Polynesian		
		Malay-Indonesian	21

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

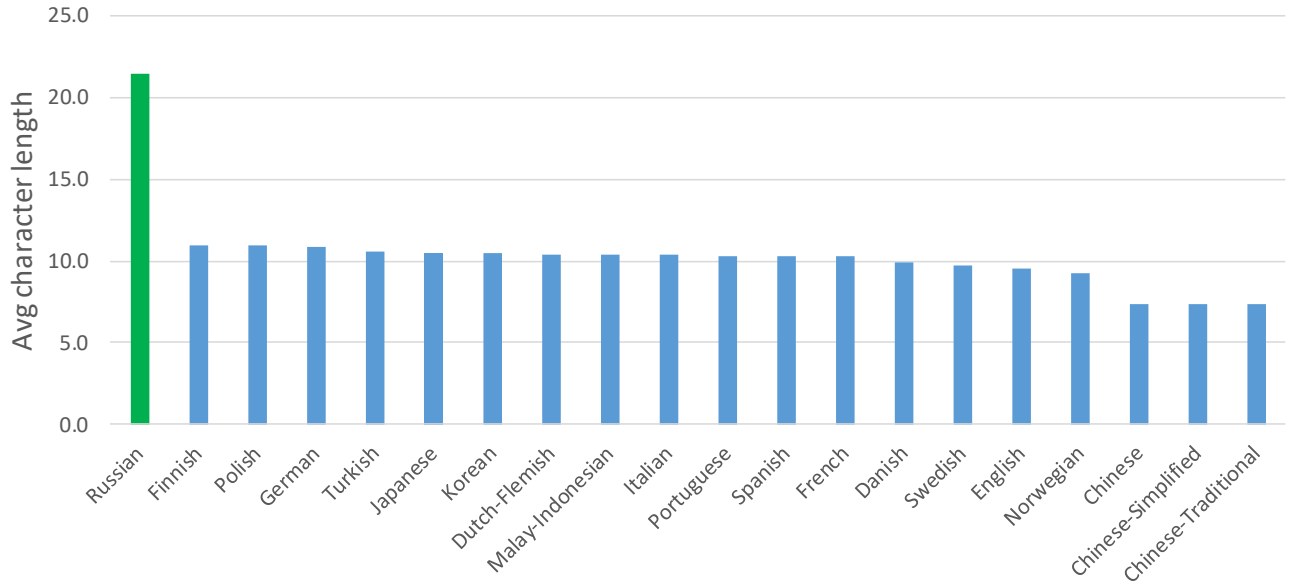
A naïve way of creating a sentiment indicator is simply to count for the number of positive (and negative) words. The positive (negative) sentiment score is the ratio of positive (negative) words over total number of words in the document. While this implementation is reasonable for English and most European languages, it would be particularly problematic for the big three Asian languages – Japanese, Chinese, and Korean. As discussed in the previous section, tokenization of these three languages tend to be quite different from English. Although we can get the native language tokenizer for Japanese, Chinese and Korean as we did for the purpose of computing language similarity factors, tokenization algorithms for many other native languages are not easily accessible.

Hence, we take an alternative approach in our sentiment analysis. We do a character string search for sentiment words in each language on raw text to get the total number of positive (negative) words in the document. We estimate the total number of words in each document using a count of total characters in the document divided by the average number of characters needed to represent a word in each language. For example, in English, the key word "insolvency" (with 10 characters) expresses a negative meaning – a company can't meet its debt payment. The equivalent word in Chinese is "破产", which comprises only two "characters".

Figure 34 shows the average number of characters for sentiment conveying words in each major language. As a benchmark, in English, to express a sentiment, on average, it takes a 9.5-character length word. On the other hand, the overall average number of letters for all English words is only 4.5. Sentiment conveying words do not contain the stop words. In NLP, stop words are filtered out before any formal analysis. Stop words are common short functional words, such as "a", "the", "in", "is", "which" and tend to have few characters in them.

As shown in Figure 34, most European languages are similar to English. Russian is one outlier – it appears that a typical sentiment word in Russian is substantially longer than in other language. On the other extreme, Chinese language seems to be rather efficient in expressing an opinion.

Figure 34 Average # of Characters of Sentiment Conveying Words, by Language

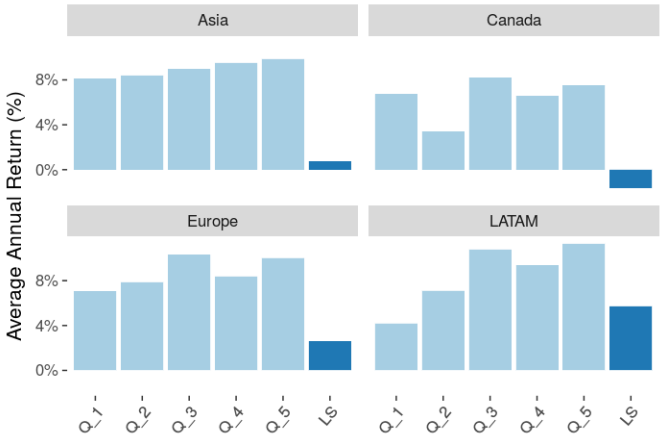


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

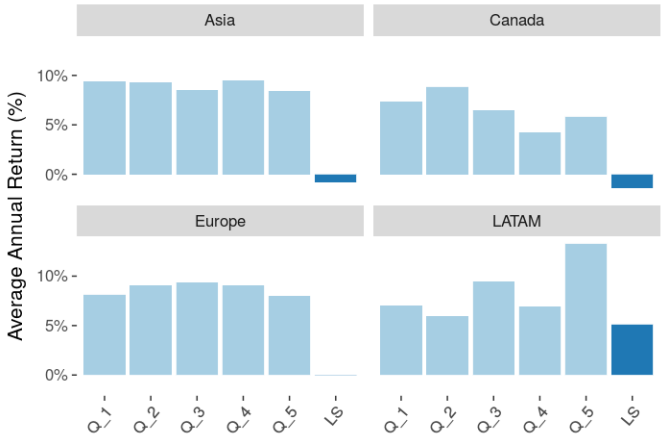
Based on annual reports, as shown in Figure 35(A), the positive sentiment factor delivers decent performance, especially in Europe and LATAM, while the negative sentiment factor struggles to add value in most regions.

Figure 35 Sentiment Measure Performance (Annual Reports)

A) Positive Sentiment



B) Negative Sentiment

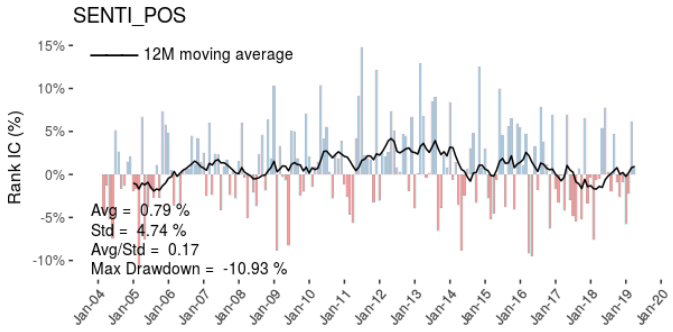


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

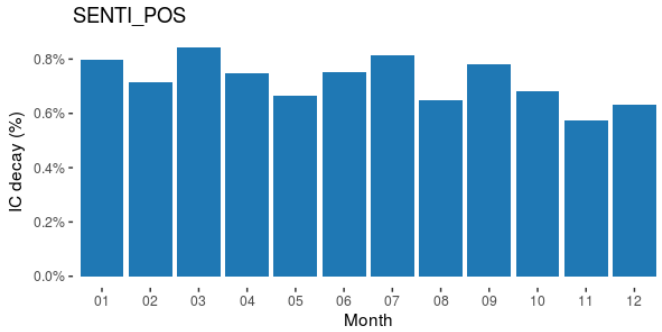
Figure 36 shows the time series performance of the positive sentiment measure in Europe and Asia. Although the average performance is modest, the information decay profile is appealing. The signal still preserves its predictive power even one year later (see Figure 36 B and D).

Figure 36 Performance of the Positive Sentiment Measure (Annual Reports)

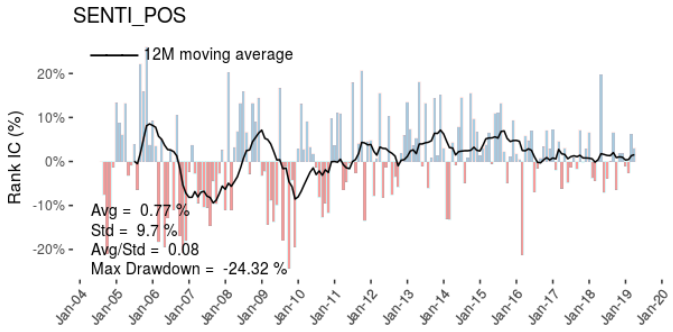
A) Rank IC (Europe)



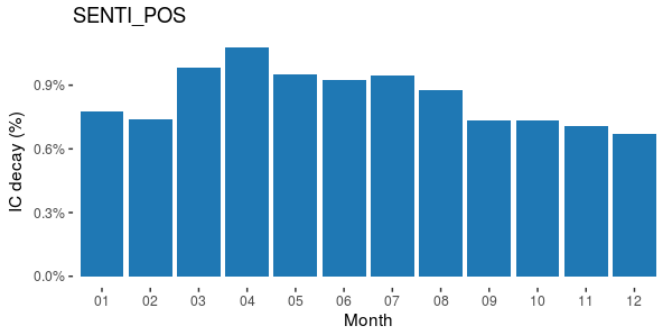
B) IC Decay (Europe)



C) Rank IC (Asia)



D) IC Decay (Asia)

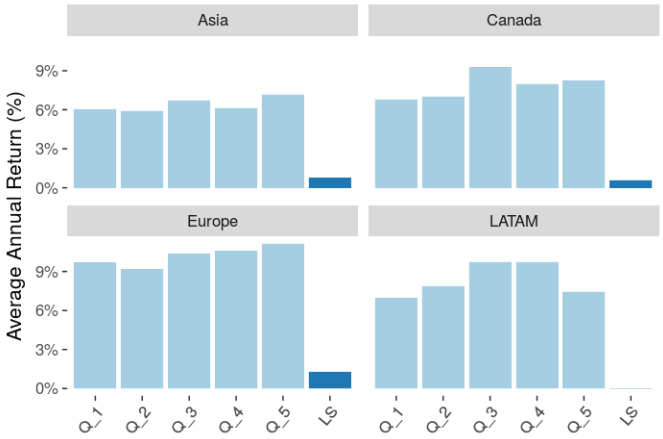


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

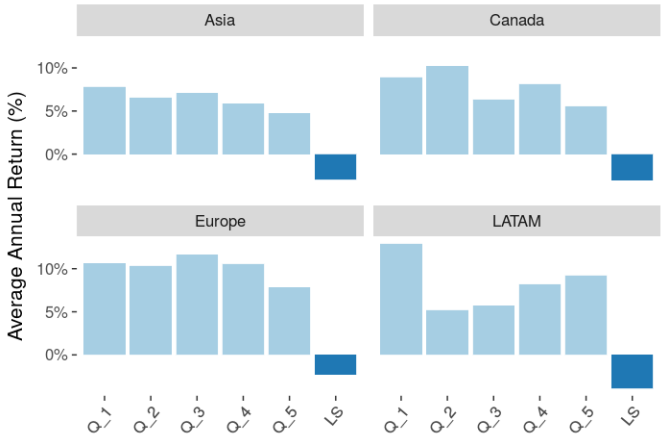
Interestingly, the sentiment factors computed on interim filings are fairly complimentary to the ones using annual reports. As shown in Figure 37, the negative sentiment factor is more useful than the positive sentiment signal, in all four regions.

Figure 37 Performance of Sentiment Measure (Interim Filings)

A) Positive Sentiment



B) Negative Sentiment

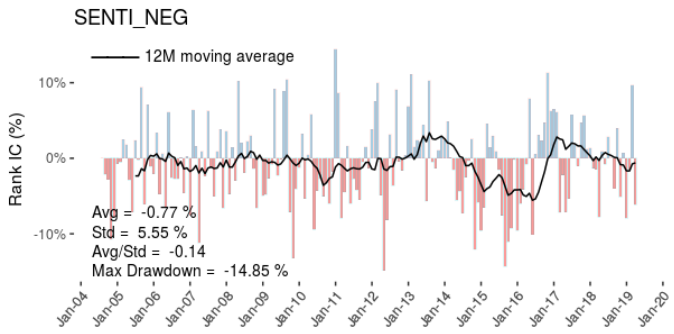


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

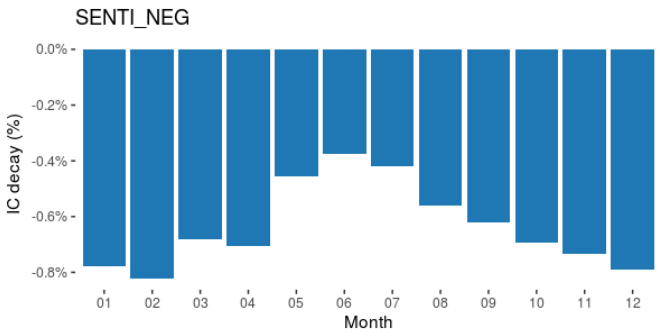
As shown in Figure 38, the negative sentiment signal is consistently negatively correlated to future stock returns in both Europe and Asia, with minimal information decay in the first 12 months.

Figure 38 Performance of the Negative Sentiment Measure (Interim Filings)

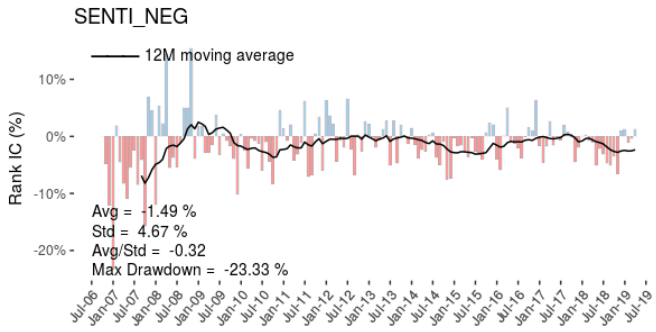
A) Rank IC (Europe)



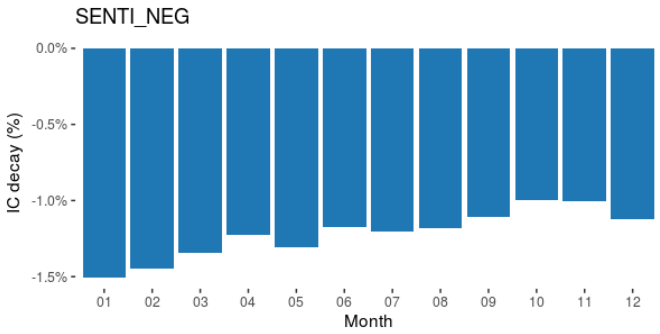
B) IC Decay (Europe)



C) Rank IC (Asia)



D) IC Decay (Asia)



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

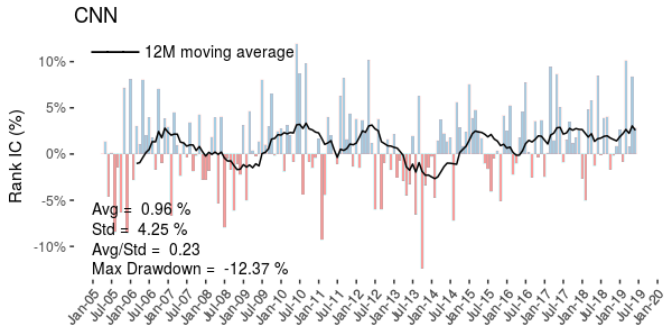
CNN – INTERNATIONAL

Similar to our CNN-US model, we apply the same CNN algorithm and structure for the four international markets.

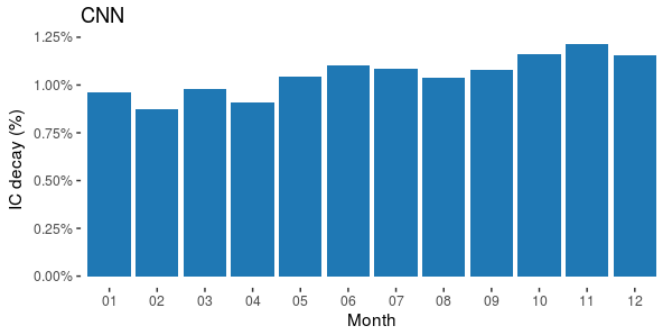
As shown in Figure 39(A) and (C), the performance of the CNN model for both Europe and Asia is slightly better than the language similarity and sentiment factors, with a similar slow decay profile (see Figure 39B and D). Unlike most conventional stock-selection factors and models that have been struggling in recent years, the CNN model is able to boost its performance as more data becomes available.

Figure 39 Performance of the CNN model (Annual Reports)

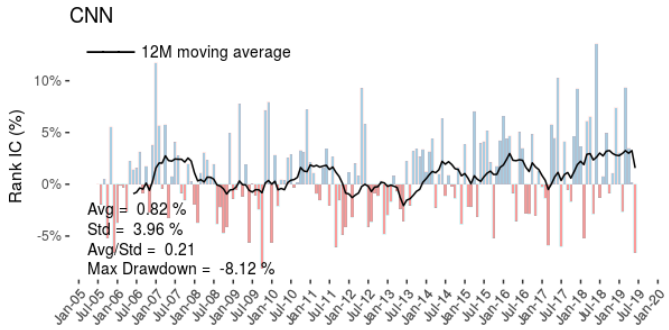
A) Rank IC (Europe)



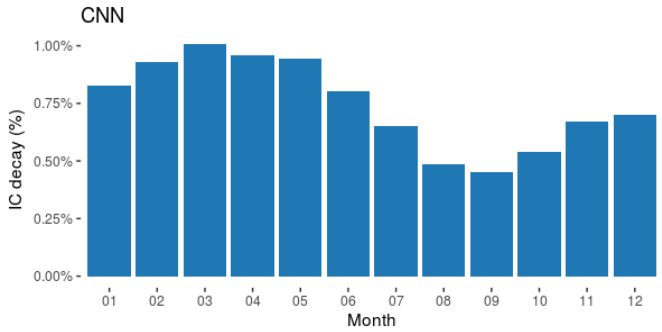
B) IC Decay (Europe)



C) Rank IC (Asia)



D) IC Decay (Asia)



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo' s QES

GINA (INTERNATIONAL) MODEL

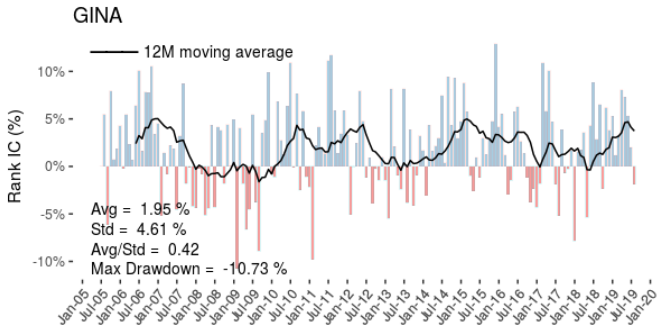
Finally, we mix the three sets of factors and models to construct our GINA-International model:

- Language similarity factors, using native language tokenization
- Sentiment measures, using our unique character-word sentiment tokenization
- Deep learning via CNN

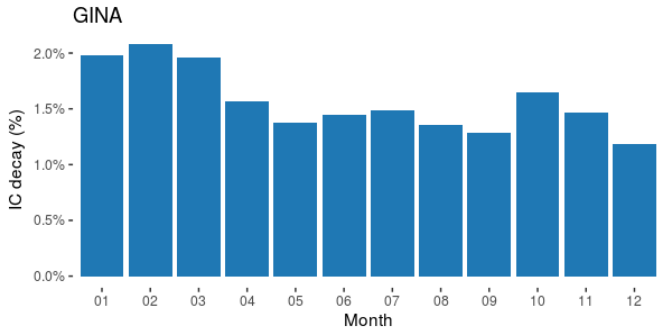
In Europe, the GINA model's performance has been reasonably consistent(see Figure 40A), with slow information decay (see Figure 40B). A long/short quintile portfolio has delivered a Sharpe ratio of 0.4x with Rank IC of almost 2%.

Figure 40 GINA Model Performance, Europe

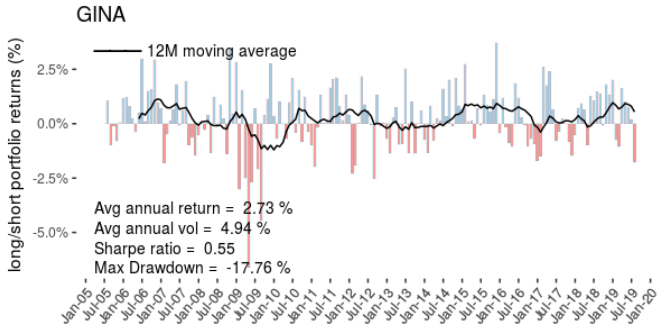
A) Rank IC



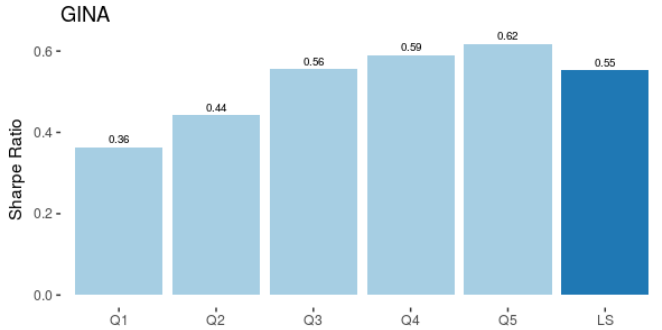
B) IC Decay



C) Long/Short Quintile Portfolio Return



D) Sharpe Ratio

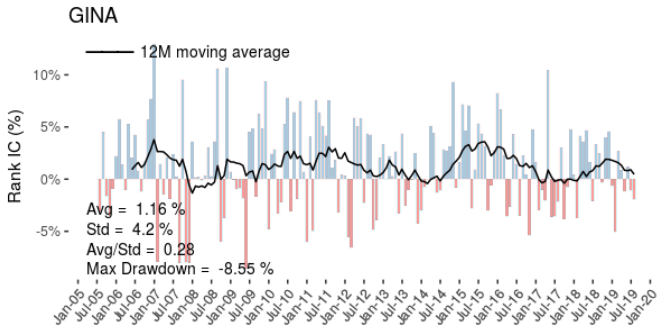


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

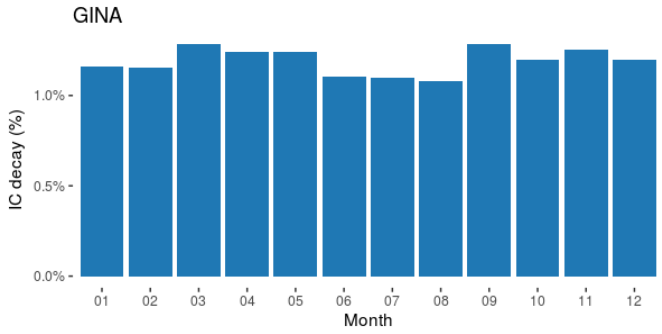
The GINA model's performance in Asia is in-line with Europe (see Figure 41), with the long/short quintile portfolio delivering a Sharpe ratio of 0.4x with Rank IC of almost 1.2%. The investment horizon is long and information decay is slow (see Figure 41B).

Figure 41 GINA Model Performance, Asia

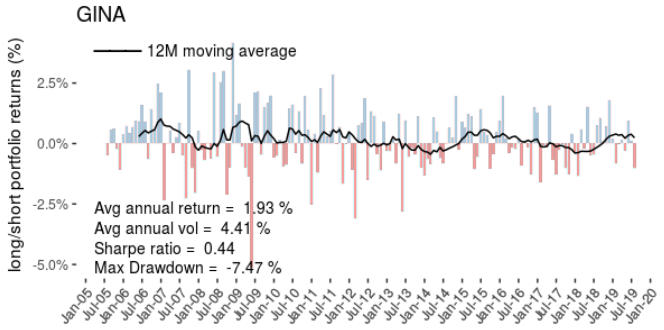
A) Rank IC



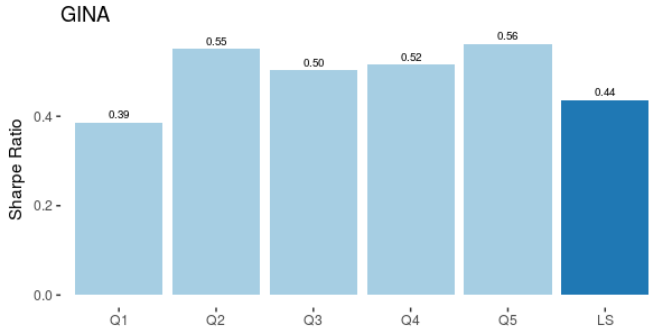
B) IC Decay



C) Long/Short Quintile Portfolio Return



D) Sharpe Ratio



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, EDGAR, Mergent, Wolfe Research Luo's QES

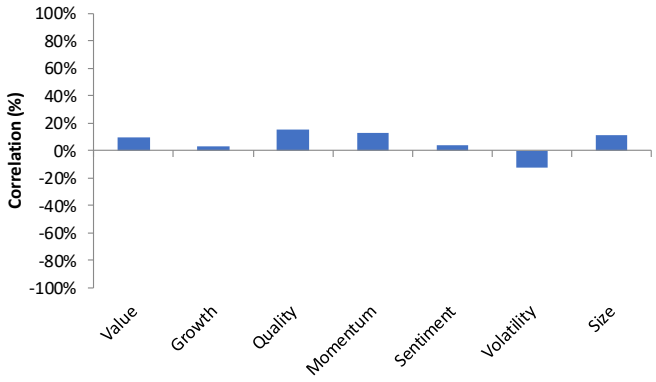
AN UNCORRELATED SOURCE OF ALPHA

Given the unique nature of the underlying data (corporate regulatory filing), the obstacles in setting up the computational technology infrastructure, and the complexity of the NLP/ML algorithms, we would expect the GINA model to be relatively uncorrelated to traditional stock-selection factors and fundamental investment strategies.

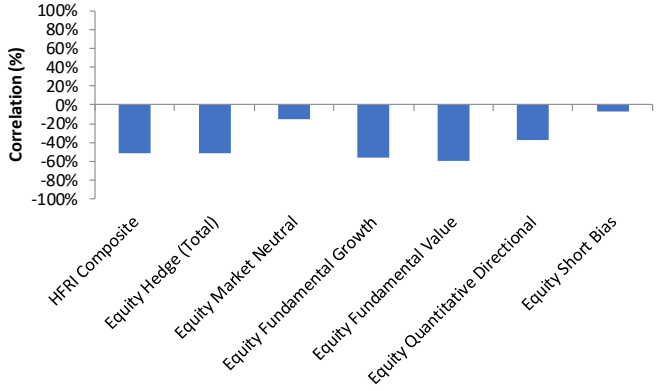
As shown in Figure 42, the GINA model's correlation with common factors is negligible. More interestingly, it is even negatively correlated to most fundamental investment styles (proxied by HFRI Indices). Therefore, we expect the GINA model to provide a considerable amount of diversification benefit to both asset manager and asset owner portfolios (see Figure 42B).

Figure 42 Correlation of the GINA Model with Common Factors and HFRI Indices, US

A) Correlation with Common Factors



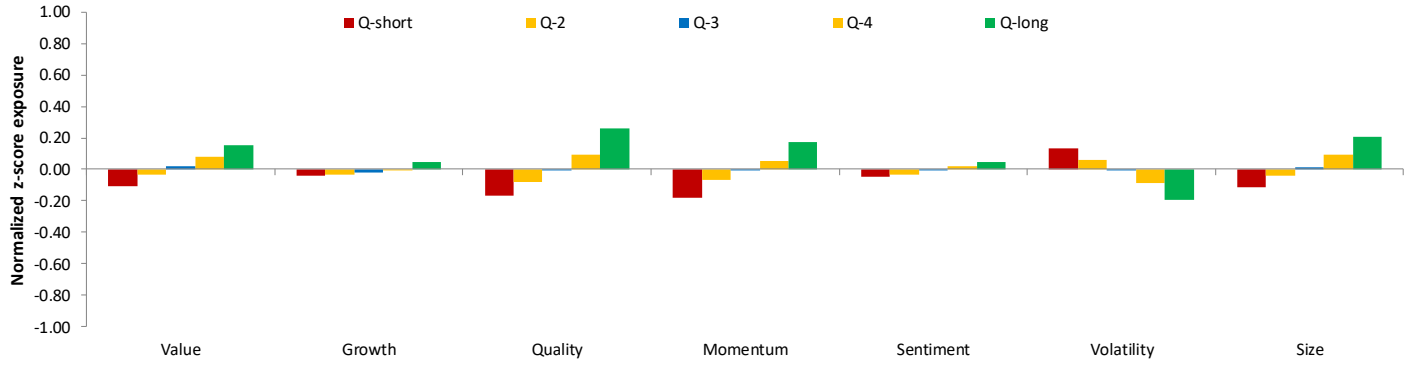
B) Correlation with HFRI Indices



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

As shown in Figure 43, the GINA signal is only marginally tilted towards valuation (cheap), high quality, low volatility, high price momentum, low volatility and large size. We want to reiterate that the overall exposures to common style factors are minimal.

Figure 43 Factor Exposure of the GINA Model, US

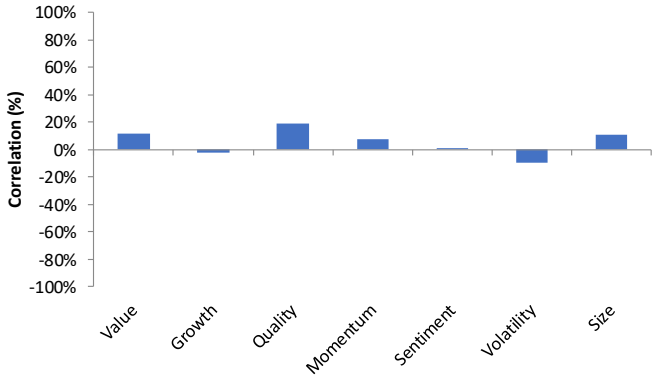


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES.

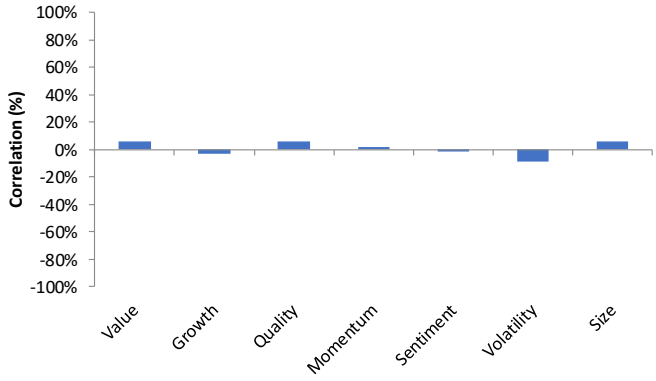
Lastly, as shown in Figure 44, the GINA model's correlation with common factors is also minimal in Europe and Asia. GINA's marginally positive exposures to value, quality, momentum and size are in Europe and Asia are roughly in-line with what we see in the US.

Figure 44 Correlation of the GINA Model with Common Factors in Europe and Asia

A) Europe



B) Asia



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

ONLINE TOOLS AND DATA FEEDS

Fundamental investors can easily access the GINA model and selected features are accessible from our online dashboard. Please note that our SPEC model dashboard is still operational, but the old SPEC model scores are also available in the new GINA tool. Given the low correlation with traditional quantitative and fundamental investment styles and the low turnover nature, the GINA model should add meaningful diversification benefit without causing significant portfolio churn.

For quantitative investors, the complete results of our backtesting are available upon request, which should assistant managers in their own research. More importantly, we offer daily data feeds that investors can plug into their own investment process directly. Given nature of the underlying textual data, the hurdle in setting up the computational infrastructure, and the sophisticated NLP/ML algorithms, we expect our models to be sufficiently different from most standard approaches on the market.

Please contact your Wolfe Research sales representative for dashboard access, model history, and ongoing data feeds.

The GINA Online Dashboard

Similar to all of our other models, we provide an easy-to-use online dashboard (see Figure 45A), where clients can run some common stock screens.

- **Filters.** On the left panel, users can choose their desired investment universe, based on the following dimensions:
    - Region – US, Canada, Europe, Asia, and LATAM
    - Country – Once a region is chosen, e.g., the Europe, users can further filter the universe by country, e.g., UK, Germany, France, etc.
    - Sector – GICS Sector classification.
- **Models and Factors.** On the right panel, users can see how each stock in the above chosen universe is ranked against other comparable firms, based on a number of proprietary models and factors.
    - GINA is our composite NLP/ML model, derived from annual and interim corporate regulatory filings. The GINA Score is in percentile, where higher scores represent higher expected alphas (i.e., expected returns).
    - SentimentML is a sentiment engine, driven by a combination of Elastic Net and xgBoost algorithms, on a set of sentiment features.
    - CNN ML is a model based on CNN (Convolutional Neural Networks), trained to predict one-year ahead stocks returns, using text from the “MD&A” and “Risk Factors” sections in the 10-K filings for US companies and the complete text extracted from the annual reports for firms outside of the US.
    - Language Similarity Annual is a YoY language similarity score, computed using the 10-K filings for US companies and annual reports for international stocks. A high score corresponds to little change in language; and therefore, is expected to have a high chance of outperformance.

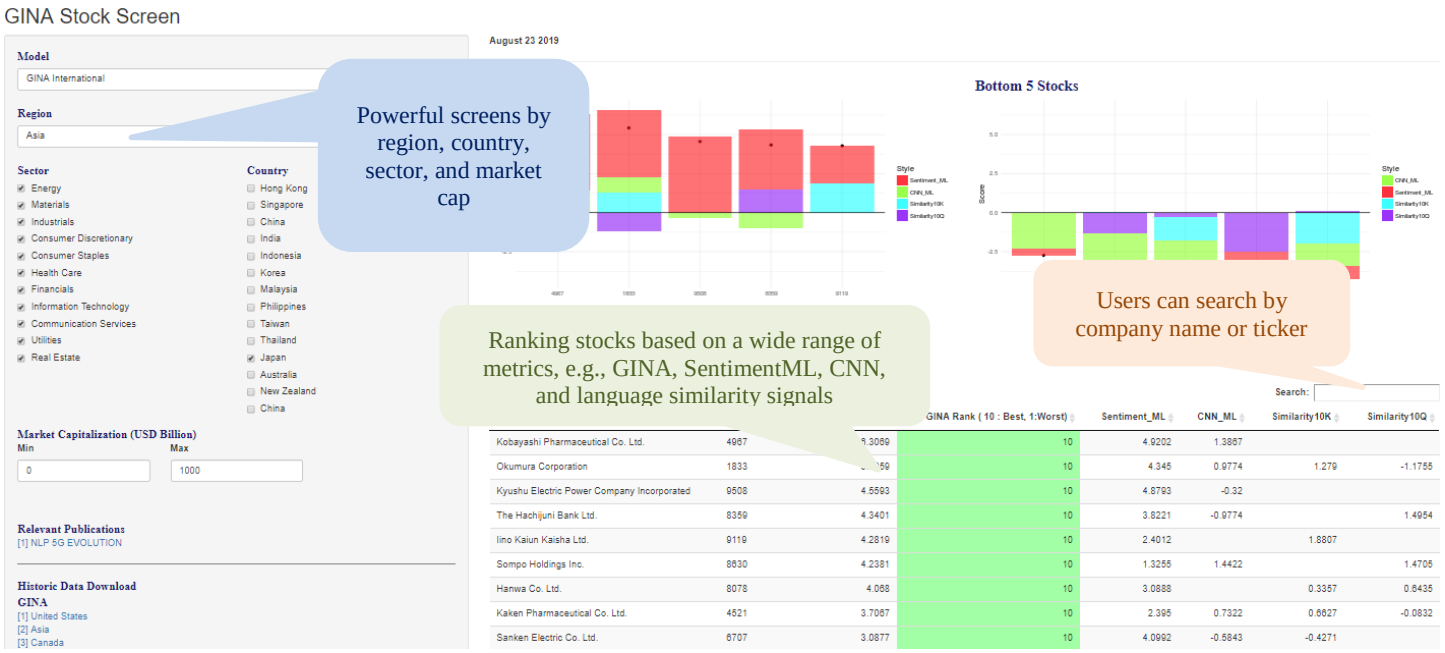
- Language Similarity Interim is a YoY language similarity score using the 10-Q (US firms) and interim filings (non-US).

- **Other Functions.** We also provide a few other common functions:
  - Search. Users can search a company of interest, by name or exchange ticker.
  - Sorting. By default, the list of stocks is sorted on the GINA model, from the best to the worse. Users can choose to rank these companies on any other metrics, either by ascending or descending order.
  - Links to Published Research. Users can download the white paper for the current paper or the previous SPEC model white paper (see *Text Mining unstructured corporate filing data*, Rohal, et al [2017]).
  - Historical Data Download. Users can download historical model/factor data for their own backtesting. Historical data is password protected. Please contact us or your Wolfe Research sales representative for the password.

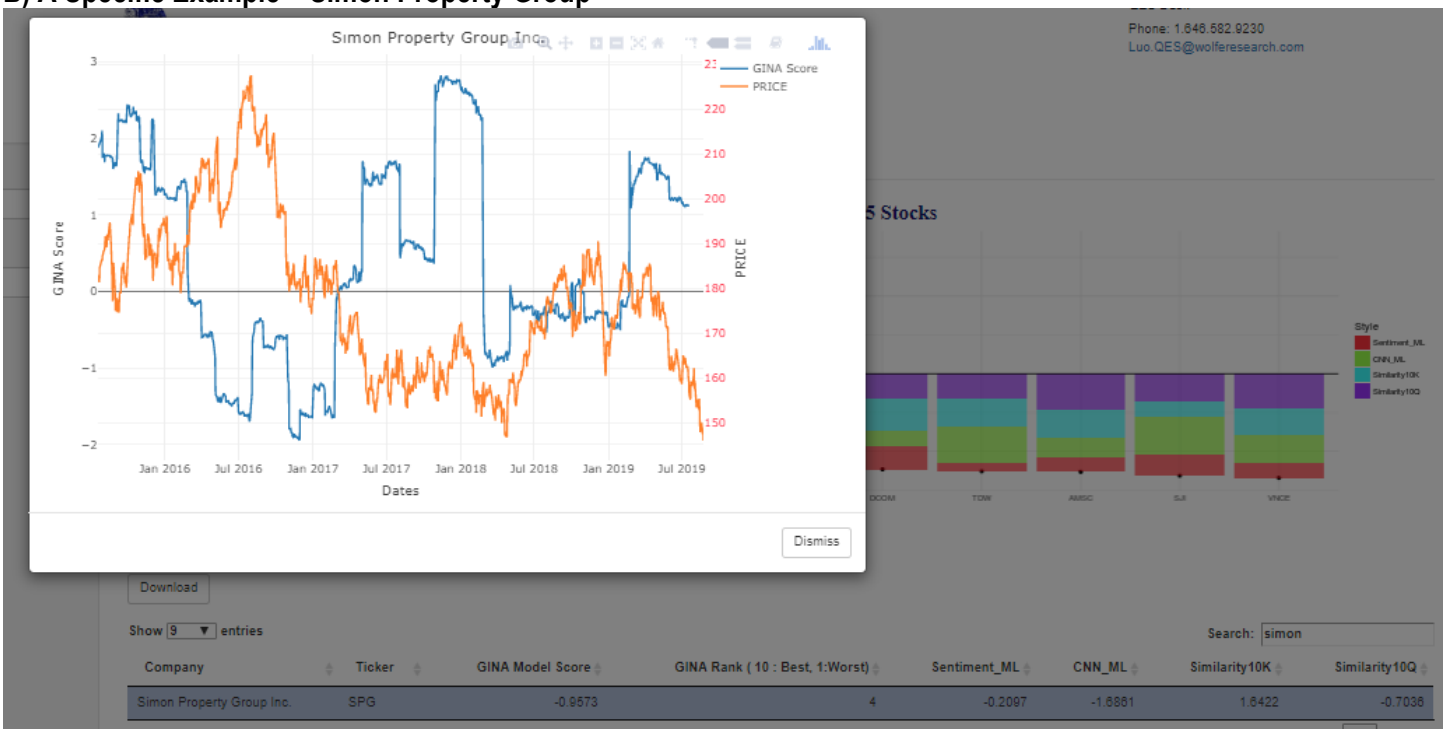
- **Single Stock Analysis.** Compared to other existing dashboards, in the new GINA online tool, we have added some functionalities for single company analysis. Users can click the name or ticker of the company of interest, the website will lead to a chart showing historical GINA rating of the stock, along with share price in the past five years.
  - Zoom and Print. Users can use the various charting tools to zoom in/out, save the chart in various formats and print.
  - A Specific Example – Simon Property Group. In Figure 45(B), we show a simple example for Simon Property Group. The blue line represents the stock's GINA score, while the orange line is share price. It is evident that the GINA score leads actual share price by around six- to nine-months. Stock price tumbles (rallies) a few months after a big change in the GINA rating.

Figure 45 GINA Online Dashboard

A) GINA Dashboard – the Landing Page



B) A Specific Example – Simon Property Group



Sources: Bloomberg Finance LLP, Factset, FTSE Russell, Markit, Ravenpack, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

Automated Data Feeds

The GINA, its underlying sub-components (e.g., SentimentML, CNN, language similarity factors) and other features are available to our clients as data feeds at various frequencies (daily, weekly, and monthly). Data feeds are typically distributed via SFTP (Secured File Transfer Protocol) or automated emails (see Figure 46).

We also provide customized stock screens for clients interested in any specific market segment.

Figure 46 Data Feed Delivery

A) FTP

B) Automated Email

Sources: Bloomberg Finance LLP, FTSE Russell, Haver, S&P Capital IQ, Thomson Reuters, Wolfe Research Luo's QES

BIBLIOGRAPHY

Ahern, K., and Sosyura, D. [2014]. "Who Writes the News? Corporate Press Releases during Merger Negotiations", Journal of Finance, http://onlinelibrary.wiley.com/doi/10.1111/jofi.12109/abstract

Amel-Zadeh, A., and Faasse, J. [2016]. "The Information Content of 10-K Narratives: Comparing MD&A and Footnote Disclosures", SSRN, https://ssrn.com/abstract=2807546

Antweiler, W., and Frank, M. [2004]. "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards", The Journal of Finance, https://www.jstor.org/stable/3694736?seq=1#page_scan_tab_contents

Ball, C., Hoberg, G., and Maksimovic, V., [2014]. "Disclosure, business change and earnings quality", SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2260371

Bodnaruk, A., Loughran, T., and McDonald, B. [2014]. "Using 10-K Text to Gauge Financial Constraints", SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2331544

Boudoukh, J., Feldman, R., Kogan, S., and Bodnaruk, M. [2013]. "Which News Moves Stock Prices? A Textual Analysis", The National Bureau of Economic Research, http://www.nber.org/papers/w18725

Brown, S., and Tucker, J. [2011]. "Large-Sample Evidence on Firms' Year-over-Year MD&A Modifications", Journal of Accounting Research, http://onlinelibrary.wiley.com/doi/10.1111/j.1475-679X.2010.00396.x/abstract

Chen, H., De, P., Hu, Y., and Hwang, B. [2013]. "Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media", SSRN, https://academic.oup.com/rfs/article-abstract/27/5/1367/1581938/Wisdom-of-Crowds-The-Value-of-Stock-Opinions

Chouliaras, A., and Grammatikos, T. [2015]. "News Flow, Web Attention and Extreme Returns in the European Financial Crisis", SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2348189

Chouliaras, A. [2015]. "The Pessimism Factor: SEC EDGAR Form 10-K Textual Analysis and Stock Returns", SSRN, http://ssrn.com/abstract=2627037

Cohen, L., Malloy, C., and Nguyen, Q. [2010]. "Lazy Prices", SSRN, http://ssrn.com/abstract=1658471

Davis, A., and Tama-Sweet, I. [2011]. "Managers' Use of Language Across Alternative Disclosure Outlets: Earnings Press Releases versus MD&A", Contemporary accounting research, http://onlinelibrary.wiley.com/doi/10.1111/j.1911-3846.2011.01125.x/abstract

Feldman, R., Govindaraj, S., Livnat, J., and Segel, B. [2010]. "Management's tone change, post earnings announcement drift and accruals", Review of Accounting Studies, 15(4), 915-953

Garcia, D., and Norli, O. [2013]. "Crawling EDGAR", The Spanish Review of Financial Economics (SRFE), http://leeds-faculty.colorado.edu/garcia/paper_edgar_v07.PDF

Hering, J. [2016]. "The Annual Report Algorithm: Retrieval of Financial Statements and Extraction of Textual Information", Academy and Industry research collaboration center, http://aircc.org/CSCP/vol7/csit76615.PDF

Hoberg, G., and Maksimovic, V. [2014]. "Redefining Financial Constraints: A Text-Based Analysis", Review of Financial studies, https://academic.oup.com/rfs/article-abstract/28/5/1312/1867105/Redefining-Financial-Constraints-A-Text-Based

Huang, A., Lehavy, R., Zang, A. Y., and Zheng, R., [2014]. "A thematic analysis of analyst information discovery and information interpretation roles", SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2409482&rec=1&srcabs=2665128&alg=1&pos=10

Hutto, C.J. and Gilbert, E. E. [2014]. "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text", Eighth International Conference on Weblogs and Social Media (ICWSM-14)

Israelsen, R. [2014]. "Tell It like It Is: Disclosed Risks and Factor Portfolios", SSRN, http://ssrn.com/abstract=2504522

Jegadeesh, N., and Wu, D. [2013]. "Word Power: A New Approach for Content Analysis", SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1787273

Jussa, J., Luo, Y., and Wang, S. [2017]. "QUANT CSI: Company Profiling Using Alternative Data", Wolfe Research Luo's QES, March 13, 2017

Lawrence, A. [2013]. "Individual investors and financial disclosure", Journal of Accounting and Economics, http://www.sciencedirect.com/science/article/pii/S0165410113000359

Li, F. [2010]. "The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach, http://onlinelibrary.wiley.com/doi/10.1111/j.1475-679X.2010.00382.x/abstract

Loughran, T., and McDonald, B. [2011]. "When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks", SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1331573

Loughran, T., and McDonald, B. [2013]. "IPO First-Day Returns, Offer Price Revisions, Volatility, and Form S-1 Language", SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2128766

Luo, Y., Jussa, J., and Wang, S. [2017a]. "The Big and The Small Sides of Big Data", Wolfe Research Luo's QES, February 8, 2017, available here

Luo, Y., Jussa, J., and Wang, S. [2017b]. "Signal Research and Multifactor Models", Wolfe Research Luo's QES, February 16, 2017, available here

Luo, Y., Jussa, J., and Wang, S. [2017c]. "Style Rotation, Machine Learning, and The Quantum LEAP" Wolfe Research Luo's QES, February 24, 2017, available here

Luo, Y., Jussa, J., and Wang, S. [2017d]. "Risk, Portfolio Construction, and Performance Attribution", Wolfe Research Luo's QES, May 9, 2017, available at here

Luo, Y., Alvarez, M., Jussa, J., Wang, S. and Rohal, G. [2018a]. "Trade war, North Korea conflict, no war after all", Wolfe Research Luo's QES, March 15, 2018, available here

Luo, Y., Rohal, G., Wang, S., Alvarez, M., Jussa, J., and Zhong, J. [2018b]. "Banking on the Banks – Welcome to BALI", Wolfe Research Luo's QES, September 4, 2018, available here

Luo, Y., Zhong, J., Alvarez, M., Jussa, J., Wang, S., Rohal, G. and Jin, Z. [2019]. "The Future of Active Management", Wolfe Research Luo's QES, May 14, 2019, available here

Mayew, W., Parsons, C., and Venkatachalam, M. [2013]. "Voice pitch and the labor market success of male chief executive officers", Journal of the Human behavior and Evolution Society, http://www.sciencedirect.com/science/article/pii/S1090513813000238

Purda, L., and Skillicorn, D. [2012]. "Accounting Variables, Deception, And A Bag of Words: Assessing the tools of Fraud Detection", SSRN, http://ssrn.com/abstract=1670832

Rohal. G., Luo, Y., Jussa, J., and Wang, S. [2017]. "Text Mining Unstructured Corporate Filing Data" Wolfe Research Luo's QES, April 20, 2017, available here

Rohal, G., Luo, Y., Alvarez, M., Jussa, J., and Wang, S. [2018]. "Tone at the Top? Quantifying Management Presentation", Wolfe Research Luo's QES, January 23, 2018, available here

Rohal, G., Luo, Y., Jussa, J., Alvarez, M., Wang, S., Zhong, J. and Jin, Z. [2019]. "Beyond Fake News", Wolfe Research Luo's QES, January 15, 2019, available here

Tetlock, P. [2007]. "Giving Content to Investor Sentiment: The Role of Media in the Stock Market", The Journal of Finance, http://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.2007.01232.x/abstract

Tetlock, P., Saar-Tsechansky, M and Macskassy, S. [2008]. "More Than Words: Quantifying Language to Measure Firms' Fundamentals", The Journal of Finance, https://www0.gsb.columbia.edu/faculty/ptetlock/papers/Tetlock_et_al_JF_08_More_Than_Words.PDF

You, H., and Zhang, X. [2007]. "Financial reporting complexity and investor under reaction to 10-K information", SSRN, http://ssrn.com/abstract=985365

DISCLOSURE SECTION

Analyst Certification:
The analyst of Wolfe Research primarily responsible for this research report whose name appears first on the front page of this research report hereby certifies that (i) the recommendations and opinions expressed in this research report accurately reflect the research analysts' personal views about the subject securities or issuers and (ii) no part of the research analysts' compensation was, is or will be directly or indirectly related to the specific recommendations or views contained in this report.

Other Disclosures:
Wolfe Research, LLC does not assign ratings of Buy, Hold or Sell to the stocks it covers. Outperform, Peer Perform and Underperform are not the respective equivalents of Buy, Hold and Sell but represent relative weightings as defined above. To satisfy regulatory requirements, Outperform has been designated to correspond with Buy, Peer Perform has been designated to correspond with Hold and Underperform has been designated to correspond with Sell.

Wolfe Research Securities and Wolfe Research, LLC have adopted the use of Wolfe Research as brand names. Wolfe Research Securities, a member of FINRA (www.finra.org) is the broker-dealer affiliate of Wolfe Research, LLC and is responsible for the contents of this material. Any analysts publishing these reports are dually employed by Wolfe Research, LLC and Wolfe Research Securities.

The content of this report is to be used solely for informational purposes and should not be regarded as an offer, or a solicitation of an offer, to buy or sell a security, financial instrument or service discussed herein. Opinions in this communication constitute the current judgment of the author as of the date and time of this report and are subject to change without notice. Information herein is believed to be reliable but Wolfe Research and its affiliates, including but not limited to Wolfe Research Securities, makes no representation that it is complete or accurate. The information provided in this communication is not designed to replace a recipient's own decision-making processes for assessing a proposed transaction or investment involving a financial instrument discussed herein. Recipients are encouraged to seek financial advice from their financial advisor regarding the appropriateness of investing in a security or financial instrument referred to in this report and should understand that statements regarding the future performance of the financial instruments or the securities referenced herein may not be realized. Past performance is not indicative of future results. This report is not intended for distribution to, or use by, any person or entity in any location where such distribution or use would be contrary to applicable law, or which would subject Wolfe Research, LLC or any affiliate to any registration requirement within such location. For additional important disclosures, please see www.wolferesearch.com/disclosures.

The views expressed in Wolfe Research, LLC research reports with regards to sectors and/or specific companies may from time to time be inconsistent with the views implied by inclusion of those sectors and companies in other Wolfe Research, LLC analysts' research reports and modeling screens. Wolfe Research communicates with clients across a variety of mediums of the clients' choosing including emails, voice blasts and electronic publication to our proprietary website.