How Have Stock Markets Responded to 35 Years of Analyst Reports? Evidence from Machine Learning and Textual Analysis*

Mustafa Gultekin† gultekin@unc.edu

Abena Owusu§ owusua@rpi.edu

Thomas Shohfi‡§ shohft@rpi.edu

Majeed Simaan¶ msimaan@stevens.edu

December 29, 2019

Abstract

Using machine learning (ML) and contrasting with simple textual sentiment score and principal components analysis (PCA) methods, we examine the time series of content within over 700,000 sell-side analyst research reports from 1983 to 2017. We find that analyst reports have significantly changed across a variety of dimensions including length and content of four existing and two new dictionaries related to valuation and alternative metrics. We find that the naive net tone of reports only explains contemporaneous, not future, equity returns. On the other hand, we find that ML methods provide substantially different results from naive sentiment and PCA approaches on determining the impact of analyst reports on financial markets. We also examine the ability of reports to predict changes in firm short interest and volatility. Overall, we find that sell-side analyst reports have stronger impact on smaller cap stocks.

Keywords: Analyst Reports, Machine Learning, Sell-Side Analysts, Textual Analysis

JEL Codes: C19, C45, C55, C63, G10, G20, G24, M41

*We thank participants at the 2019 Eastern Finance Association Annual Meeting and R/Finance 2019. We also thank seminar participants at Rensselaer Polytechnic Institute and Stevens Institute of Technology. We thank the Donald Shohfi Financial Research Fund for computing support. We thank Yuvraj Chopra and Avnish Grover for research assistance. All errors are our own.

†Kenan-Flagler Business School, University of North Carolina, 300 Kenan Center Drive, Chapel Hill, NC 27599, USA.

‡Corresponding author. Tele: (518) 276-6582, Fax: (518) 276-8661, Web: shohfi.com

§Lally School of Management, Rensselaer Polytechnic Institute, 110 8th Street, Pittsburgh Building, Troy, NY 12180, USA.

¶School of Business, Stevens Institute of Technology, 1 Castle Point Terrace, Hoboken, NJ 07030, USA.

How Have Stock Markets Responded to 35 Years of Analyst Reports? Evidence from Machine Learning and Textual Analysis

Abstract

Using machine learning (ML) and contrasting with simple textual sentiment score and principal components analysis (PCA) methods, we examine the time series of content within over 700,000 sell-side analyst research reports from 1983 to 2017. We find that analyst reports have significantly changed across a variety of dimensions including length and content of four existing and two new dictionaries related to valuation and alternative metrics. We find that the naive net tone of reports only explains contemporaneous, not future, equity returns. On the other hand, we find that ML methods provide substantially different results from naive sentiment and PCA approaches on determining the impact of analyst reports on financial markets. We also examine the ability of reports to predict changes in firm short interest and volatility. Overall, we find that sell-side analyst reports have stronger impact on smaller cap stocks.

Keywords: Analyst Reports, Machine Learning, Sell-Side Analysts, Textual Analysis

JEL Codes: C19, C45, C55, C63, G10, G20, G24, M41

1 Introduction

Analyst reports contain important new information and provide interpretation of existing information that elicits market reactions (Asquith et al., 2005) and can enhance capital market functionality (Merkley et al., 2017). Typically the informativeness of analyst report content is characterized by quantitative information including analyst recommendations, price targets, valuation models, and earning forecasts. Nonetheless, these reports also contain significant textual content that investors further comprehend along with quantitative content.

While early studies of analyst report content, including Previts et al. (1994) were limited by computational constraints, more recent studies have examined information content (Frankel et al., 2006; Huang et al., 2014), complementation with corporate disclosures (Chen et al., 2010; Frankel et al., 2006), and readability (De Franco et al., 2015; Lehavy et al., 2011). For instance, Huang et al. (2014) analyze the textual content of more than 350,000 analyst reports for S&P 500 member firms from 1996 to 2008. The authors use sentence-level Bayesian classification to determine sentiment and show that analyst report content has become less optimistic after 2001. They also find that the marketplace reacts more strongly when analyst report content conveys more bad news.

In our paper, we use a large sample of more than 700,000 sell-side analyst reports covering almost 1,000 firms from 1983 to 2017. Over the sample period, we find that the average number of pages per analyst report has more than doubled, increasing from 6 pages to 14 pages per report. This is also reflected by the number of words that has more than tripled. Given this large corpora of analyst reports, we apply both machine learning and textual analysis methods to improve extraction of information and examine the value of this information to market participants. Specifically, we examine the ability of textual measures constructed using simple sentiment scores, principal components analysis, and machine learning in predicting monthly abnormal returns, short interest and realized volatility.

We set our investigation by first examining the time series of content within analyst

reports. We find the length of analyst reports has steadily increased and that aggregate sentiment exhibits strong annual variation. Reports have been more negative and uncertain since the 2003 Global Research Settlement with more litigious and less positive content in recent years. We introduce two new dictionaries: one based on valuation terms and another based on alternative data. We find that alternative data terms were used most during the late 1990s technology bubble. Our time series data provide new evidence of changing analyst behavior over time expressed in the content of analyst reports.

Our second set of analyses deploys both unsupervised and supervised textual scores at the firm-month level. To assess the applicability of these methods, we run a non-parametric test using portfolio formation. At each month, we rank firms based on a representative textual content index and investigate the next month's return. In addition, we also investigate the contemporaneous return of each portfolio. When using a conventional net tone sentiment score, as in Huang et al. (2014), we observe that analyst reports have no value in explaining future month returns. At the same time, we find consistent evidence indicating that analyst sentiment is more driven by current returns rather than the other way around. For instance, by ranking stocks into terciles based on net tone, we find that a value-weighted portfolio with the highest net tone exhibits past outperformance of 8% per annum over a value weighted portfolio consisting of stocks ranked as expressing lower net tone. However, returns of the next month within each portfolio are relatively similar (i.e. zero alpha). Overall, we find that the findings are sensitive to the choice of sentiment index. Conventional net tone sentiment, which is commonly used in the literature, appears to misrepresent analyst content and serves less to explain the relationship textual content of these reports and market reaction.

While such evidence may indicate that analyst sentiment is driven by the market rather than the other way around, we should take into consideration limitations behind using this approach. In the net tone approach, one assigns equal weights to positive and negative terms. However, as Jegadeesh and Wu (2013) indicate, the weights assigned to terms (or dictionaries in our case) is critical. Specifically, they state that the "appropriate choice of term weighting in content analysis is at least as important as, and perhaps

more important than, a complete and accurate compilation of the word list." Motivated by this, we repeat the same analysis but using different weighting schemes. By running principal components analysis (PCA) across six different textual measures on a rolling window, we show improvements in value-weighted following month portfolios yielding around 90 bps more risk-adjusted per annum than naive sentiment (but remain statistically insignificant). When we consider PCA and contemporaneous returns, we find no evidence to claim that analysts' textual content is driven by current stock performance.

The PCA, nonetheless, is an unsupervised data driven method that combines the textual content of analyst reports in a single index. However, such an index lacks market feedback as in Jegadeesh and Wu (2013) to find the optimal weights allocated to different dictionaries. To address this, we use an elastic net least squared regression with 10 folds cross-validation. The algorithm is efficiently executed on a rolling window, with respect to the package developed and maintained by Friedman et al. (2010).[1] At each month, we use an estimation window of five years to estimate the abnormal return of each stock covered by the analysts. Given the computed alphas and the elastic net algorithm, we determine the loadings (weights) of each dictionary. Finally, we construct a firm-month supervised textual index, which we refer to as the ML score.

To test the implications of the ML score, we repeat the same non-parametric test to investigate the market reaction to the covered stocks. In line with PCA results, we find that the ML score provides some forward looking value. In particular, a long-short strategy that goes long (short) a value-weighted portfolio of stocks ranked in the highest (lowest) ML score yields 80 bps more risk-adjusted annual return than naive sentiment, after controlling for Fama and French (1993) and Carhart (1997) risk factors. Nonetheless, this evidence is sensitive to firm size. For small cap stocks, which we identify as firms with total book value of assets below the cross-sectional median at each quarter, we find opposite evidence. In particular, we find that larger ML score is associated with lower future risk-adjusted returns. For instance, an equal-weighted portfolio that goes long and

[1]We also repeat the results using a support vector machines (SVM) with a linear kernel. Running an SVM model with the kernel trick is equivalent to running a neural network with a single hidden layer. We find similar results using either approach, however, we report the elastic net algorithm since it less opaque than the SVM.

short in the small cap stocks that fall in the highest and lowest ML score, respectively, returns a negative risk-adjusted return of 2.8% per annum, which is significant at the 5% level.

We subsequently examine different machine learning methods on the prediction of firm short interest and volatility. We focus on short interest, since the percentage of shares that investors short sell can be an indicator of the market's sentiment of whether to expect a fall in value of the shares. Existing literature also find a negative significant relationship between short interest and stock returns (see e.g., Asquith and Meulbroek (1995); Desai et al. (2002)). Engelberg et al. (2012) show that short sellers successfully process negative information in their trading strategies. Hence, we investigate whether the the textual content within the analyst reports also predict short selling. We find that net tone is positive and significant at the 1% level for the contemporaneous and the next quarter short interest. This indicates that more positive tone relative to negative is associated with lower short selling in the current and next quarters. Put differently, more negative tone is associated with more short selling.

On the other hand, we find a positive significant (no) relationship between the ML (PCA) score and current/next quarter short interest. To better understand this result, we note that the ML score is extracted with respect to future abnormal returns. In addition, Rapach et al. (2016) show that short interest is a strong predictor of stock returns and that a one-standard-deviation increase in the aggregate short interest level corresponds to a 6% decrease in the future annualized market excess return. Hence, combined with the findings from the non-parametric test, we expect that the ML score to have a negative (positive) loading on current/future short interest for large (small) stocks. Specifically, we find the panel data evidence is mainly evident across small cap stocks, however, insignificant among large cap stocks.

Finally, we extend our analysis to stock price volatility. We find that only ML scores explain current and subsequent quarter volatility. Small cap stock volatility has slightly stronger reactions to analyst reports but the effect is strongest in the aggregate sample.

Our findings, overall, indicate that there is a different market reaction to analyst re-

ports for small and large cap stocks. This evidence is consistent with small cap stocks having a poorer information environment that results from analyst coverage decisions (Brennan and Hughes (1991)) and differing market reactions to information flows (Hong et al. (2000)). In addition, the ML findings stress the importance of using further advanced tools to extract relevant textual information. On one hand, naive net tone approach is transparent and easier to comprehend. On the other hand, it is critical to find optimal words for different terms to better capture the semantics of a textual corpus. While our analysis is limited to dictionaries use only, the data-driven approach can be extended to analyze specific words and deploy more advanced tools such as word2vec (see e.g., Mikolov et al. (2013)).

The rest of the paper is organized as follows. In Section 2, we provide an overview the data sources used along with a number of summaries. Section 3 covers the non-parametric tests along with the extraction of the data-driven firm-month indices. We dedicate Section 4 to the parametric tests, in which we conduct a series of panel regressions using the extracted textual scores. Finally, Section 5 concludes.

2 Data Description and Summary

In this section, we provide details regarding the data used in this paper as well as a summary of the constructed textual data.

2.1 Data Sources

We integrate multiple sources of data to implement our empirical investigation. For the textual data, our main data of interest are derived from sell-side analyst reports obtained from Investext. We download reports in portable document format (PDF) and convert each report to text using the application pdftotext.[2] If the PDF document does not contain internally searchable text (i.e. the PDF is image based), we use Tesseract open source optical character recognition (OCR) software to extract text from each an-

[2]pdftotext is part of the open source project Xpdf available at http://www.xpdfreader.com/

alyst report.[3] Investext also provides summary information for each analyst report file including report title/subtitle, date of release, brokerage name, analyst name, and language. We limit this initial sample only to English language reports. The initial sample includes approximately 950,000 analyst reports with unique identification numbers covering more than 1,500 randomly selected U.S. exchange listed firms from 1982 through 2018. These PDF (converted text) sell-side analyst reports are more than 304 (30) gigabytes in size.

For quantitative data, we refer to CRSP for stock returns and market data that is used to construct other control variables (e.g. Tobin's Q, etc.). Short interest and firm financial report data are from Compustat. In each report, we extract the NCUSIP from each report in order to merge the textual data with other sources. For CRSP, for instance, we keep data points that correspond to a unique NCUSIP-PERMCO identification. This reduces the number of unique firms to 1,270.

In the CRSP data set, we focus on common shares alone, i.e. first digit of the share code is equal to one. Moreover, we remove any missing values for prices and returns. We compute the market capitalization for each firm by multiplying the share price by total shares outstanding. Additionally, we retain firms that have at least 60 months of stock returns. We do so in order to have sufficient sample size to construct the proposed ML data-driven sentiment index, which we will discuss later on. These filters produce our final sample which includes 724,829 reports covering 843 firms.

2.2 Textual Analysis

For each underlying firm, there are multiple analyst reports released over coverage time periods. For instance, for firm i at the end of month t, there are A analyst reports covering the company in that specific month. For each analyst report, a number of sentiment variables are extracted in line with Loughran and McDonald (2011). Additionally, we consider other textual content. Brown et al. (2015) find that sell-side analysts rely on valuation models to support their recommendations therefore we examine valuation

[3]Tesseract is available at https://github.com/tesseract-ocr/. Older analyst reports in Investext (i.e. prior to 2000) are more frequently image-based and require OCR for textual analysis.

related terms (e.g. discounted cash, dividend discount, and etc.). Moreover, when traditional valuation metrics are not applicable or sufficient, we also consider alternative metric terms. Analysts that incorporate information beyond traditional financial market and accounting metrics may have a disparate impact on markets relative to peers. Painter (2018), for example, find that investors with relevant satellite imagery data (e.g. parking lot traffic) possess an advantage over other market participants. We list all words for these two additional dictionaries in Table 1.

Table 1: **Valuation and Alternative Terms**

Dictionary	Terms
Valuation	cf, discounted, cash, discount, cash, valuati, sum, of, the, part, sum, of, part, dividend discount, comparables, comparable, firm, comparable, compan, liquidation, val, replacement cost, peer, group, ratio, multiple, price-to-, price, to, present, value
Alternative	channel, check, data, satellite, imag, geolocation, sensors, social, media, web search, parking, lot, traffic, sentiment, textual, logistics, biometric

We note that while the release of analyst reports takes place on a frequent basis, the timing of these reports is not constant over time. For this reason, for each firm at a specific month, we aggregate the textual content on the firm-month level by taking the average. In the event that there were no releases in a single month, we use a decay factor of 50%, such that textual content of the current month constitutes 50% of the previous month. If in the next month there are no releases, then we construct the textual data by taking 0.5 of the previous content, such that the original release now weighs 0.25. We repeat this procedure for each firm-month until a new analyst report is released. Hence, this allows us to expand the textual data to a greater horizon, while at the same time, controlling for information decay.

2.3 Summary of Textual Content

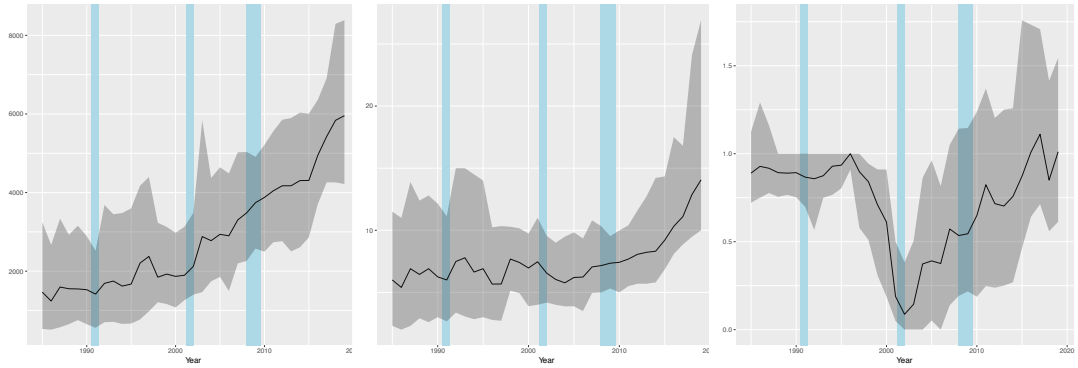
Our first sentiment measure is constructed as net sentiment tone, which corresponds to the difference between the number of positive and negative terms defined by Loughran and McDonald (2011) adjusted by the total number of words. In particular, we map each

report covering company i at the end of month t by analyst a into a sentiment score as

$$Z_{i,t,a} = \frac{N^{+}_{i,t,a} - N^{-}_{i,t,a}}{N_{i,t,a}} \tag{2.1}$$

where $N^{+}_{i,t,a}$ ($N^{-}_{i,t,a}$) denotes the number of positive (negative) words in a single report covering company i at the end of month t by analyst a. Moreover, $N_{i,t,a}$ denotes the number of words in the that specific report. In addition, to extract a single sentiment index for each firm i at the end of month t, we compute the mean of the statistic in (2.1), $Z_{i,t}$.

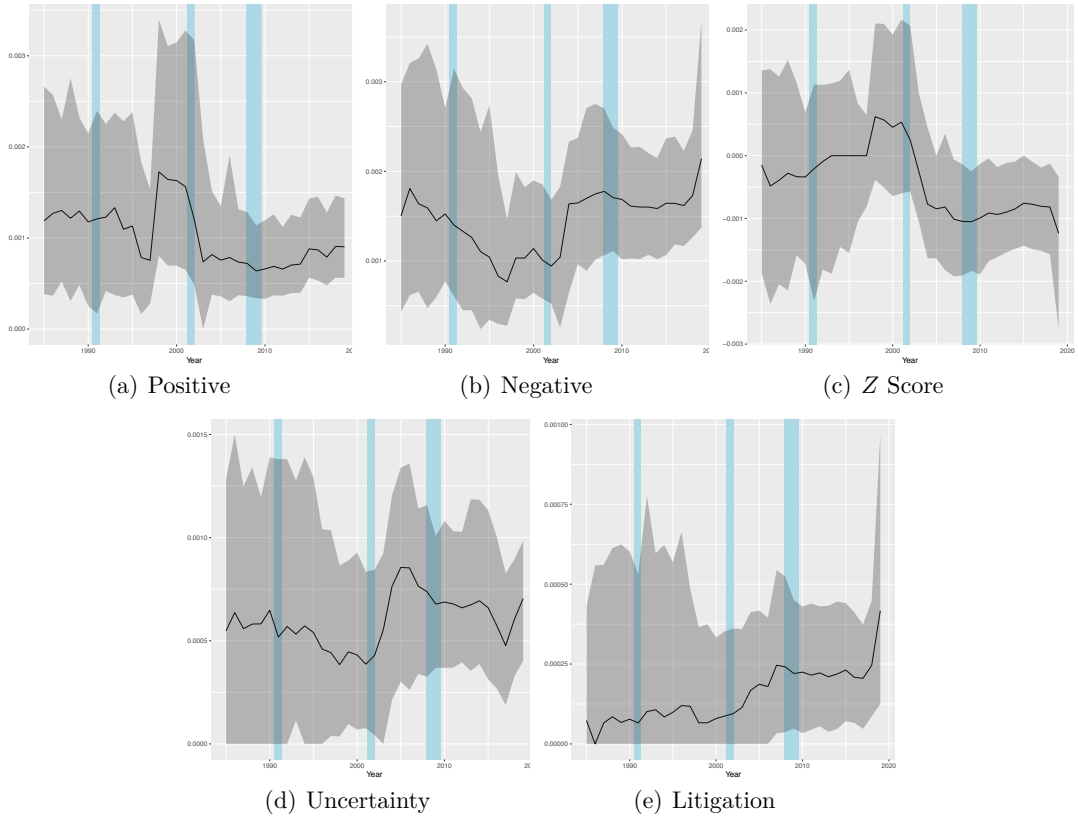
Figure 1: **Time Series of Analyst Report Characteristics**
The following panels demonstrate the distribution of analyst report length (panels (a), (b), and (c))



(a) Number of Words (b) Number of Pages (c) Number of Images per Page

In Figure 1, we demonstrate the mean distribution of the report characteristics over time. In particular, Panels (a) and (b) report the distribution of the average number of reports per firm as well as the number of pages per company covered in a single year. For each year, we highlight the 10%, 50%, and the 90% percentiles. In either case, we note that report length and number of reports has increased over time. At the same time, we note that there is heterogeneity in the data in terms of the spread between the 90% and the 10% percentiles. Panel (c) displays the number of images extracted from each analyst report page. Graphical data within sell-side reports exhibit a different trend than textual data. The number of images per page was higher before 1995 and fell precipitously during the late 1990s tech boom only to recover slowly over the next twenty years.

Figure 2: **Loughran and McDonald (2011) Dictionary Ratio**
The following panels demonstrate the distribution of the mean sentiment given the dictionaries proposed by Loughran and McDonald (2011). Each panel reports the number of sentiment dictionary over the total number of words in the report.



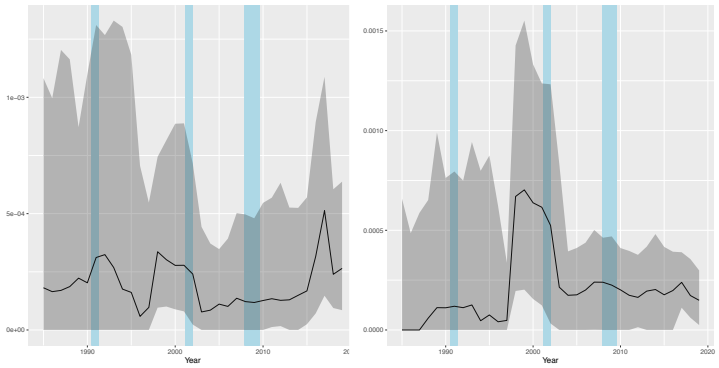
(a) Positive (b) Negative (c) Z Score

(d) Uncertainty (e) Litigation

In Figure 2, we illustrate four Loughran and McDonald (2011) dictionaries over time. In Panel (a) and (b), we plot the positive and negative ratio, respectively. We observe that positive and negative tone exhibit inverse relation. For instance, there is an increase in positive tone around 2000 with a simultaneous decrease in negative trend. Examining the net tone in Panel (c), i.e. the score described in (2.1), we note that, overall, there is a decline in net tone over time. Nonetheless, we also observe that there is an increase in net sentiment leading up to the dot com bubble. At the same time, we also observe a relative increase in the net tone post the recent financial crisis. Looking at uncertainty related terms, we observe in Panel (d) an increase in uncertainty surrounding the 2007-09 financial crisis. On the other hand, litigation related terms seem to be on an increase over time, as Panel (e) demonstrates.

Figure 3 displays time series variation for the valuation and alternative dictionaries. Valuation related content exhibits three spikes and subsequent reversals in the early 1990s, late 1990s, and late 2010s, respectively. Alternative metric content shows a major spike in the late 1990s with a correction, and subsequently higher period of mean content, following the tech bubble burst.

Figure 3: **Additional Dictionaries**
The following panels demonstrate the distribution the mean ratio of two specific dictionaries. Panel (a) refers to valuation-related terms and Panel (b) refers to alternative dictionary terms, see Table 1 for further information.



(a) Valuation (b) Alternative

3 Portfolio Analysis

In this section, we conduct a battery of non-parametric tests to investigate the market reaction to textual content of analyst reports. We set our investigation by first using the naive sentiment approach to sort firms into portfolios. Given the baseline results, we refer to more advanced data driven approaches to extract a representative textual score for each firm-month in the data. Given each, we repeat the same portfolio formation analysis.

3.1 Baseline Results

At the end of each month, we have a sentiment measure for each company. To form portfolios, we group companies into quartiles based on the mean sentiment score $Z_{i,t}$.

Firms with highest (lowest) sentiment score are ranked in the top (bottom) tercile. If the current month doesn't have a sentiment score (i.e. no reports were available) then we group the firm based on the previous sentiment score. This is repeated until a new sentiment score shows up. Finally, after merging, the report data with the CRSP, we end up with 456 firms and 407 months, totaling 136,911 firm-month observations.[4]

In Table 2, we provide summary statistics for the number of firms in each tercile. We note that are around roughly 60 firms in each portfolio on average. We also note that while on average we have a comparable number of firms in each tercile, the distribution varies across the four portfolios.

Table 2: **Distribution of Firms across Portfolio Terciles**

Portfolio	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
1	414	177.993	95.004	13	78	261	310
2	414	173.150	95.057	11	76.5	260	309
3	414	172.481	95.176	11	76.2	259.8	307

In Panel (a) of Table 3, we report the OLS regression results for the realized portfolio returns over the next month realized return. We adjust the return of each portfolio with respect to Fama and French (1993) and Carhart (1997) risk factors. We scale the risk-adjusted returns, i.e. the alphas, by 12 to report them in annual basis. At the same time, we report the coefficient for each risk factor.

A number of comments are in order. First we note that there is a slight monotonic increase in the risk-adjusted return (as well as raw returns) as the net tone increases. Nonetheless, the difference between the 3rd and 1st terciles is around 1% and statistically insignificant. Second, it appears firms associated with low sentiment scores have higher value premiums than those with high sentiment scores as the HML coefficient illustrates. This could suggest that analysts seem more optimistic about growth stocks than value stocks. Nonetheless, such result is insignificant. Third, we observe that firms with high sentiment scores do exhibit positive momentum. This implies that analysts' sentiment

[4]Note that the increase in the whole sample is due to the fact that we extend the sentiment from previous periods. This creates some duplication in the data, however, allows us to construct portfolios on a more consistent level.

Table 3: **Portfolio Sorting - Value Weighted**
This table presents estimates from OLS regression of monthly value-weighted excess returns on each sentiment-sorted portfolio on the three Fama and French (1993) and Carhart (1997) momentum risk factors. Portfolios are sorted on a monthly basis with respect to the median net tone described from (2.1). Alphas are reported on an annual basis, and standard errors are adjusted for heteroskedasticity and autocorrelation using Newey and West (1986) with three lags.

	Panel A, Dependent variable: next month excess return			
	(1)	(2)	(3)	(3)-(1)
Alpha	0.016	0.017	0.030***	0.014
	(0.010)	(0.011)	(0.011)	(0.015)
MKT	0.967***	1.000***	1.030***	0.063**
	(0.023)	(0.026)	(0.026)	(0.030)
SMB	−0.157***	−0.195***	−0.143***	0.014
	(0.029)	(0.067)	(0.038)	(0.038)
HML	0.107***	0.026	0.023	−0.084
	(0.033)	(0.039)	(0.048)	(0.056)
MOM	−0.093***	0.070**	0.064***	0.157***
	(0.024)	(0.030)	(0.022)	(0.035)
Observations	414	414	414	414
R^2	0.871	0.831	0.844	0.092
Mean	0.127	0.139	0.155	
Std	0.156	0.160	0.164	
SR	0.815	0.867	0.942	
	Panel B, Dependent variable: current month excess return			
	(1)	(2)	(3)	(3)-(1)
Alpha	0.027***	0.100***	0.110***	0.082***
	(0.009)	(0.015)	(0.014)	(0.018)
MKT	0.989***	0.985***	1.004***	0.015
	(0.021)	(0.028)	(0.029)	(0.036)
SMB	−0.216***	−0.125***	−0.059	0.157***
	(0.032)	(0.039)	(0.042)	(0.057)
HML	0.095**	−0.013	0.018	−0.077
	(0.039)	(0.051)	(0.053)	(0.062)
MOM	−0.061*	0.036	0.108***	0.168***
	(0.032)	(0.034)	(0.032)	(0.057)
Observations	414	414	414	414
R^2	0.880	0.808	0.817	0.123
Mean	0.140	0.216	0.234	
Std	0.156	0.162	0.163	
SR	0.893	1.337	1.432	
Note:				*p<0.1; **p<0.05; ***p<0.01

may be driven by previous performance.

To gain a closer perspective on the momentum effect, we consider the following test. Rather than computing the next month risk-adjusted return, we compute the current risk-adjusted return of the portfolios. This is an in-sample test to investigate the degree of which analyst' sentiment is driven by the former performance of stocks. Similar to Panel

(a) of Table 3, we report the OLS regression results for the current portfolio returns in Panel (b). We observe that there is a monotonic increase in the risk-adjusted returns, whereas the difference between the top and the bottom terciles yields a risk adjusted return of 8.2% significant at the 1% level. Similarly, looking at the momentum effect, we observe that the current returns of the top sentiment portfolio exhibit higher momentum than the bottom sentiment portfolio.

The evidence from Table 3 indicates that analyst' sentiment is largely driven by simultaneously realized stock performance. At the same time, we observe the tone of analyst reports' has insignificant impact on the next month returns. As an untabulated additional check, we repeat the analysis from Table 3 by considering an equal-weighting scheme. Similar to the former, we find consistent evidence suggesting that the analyst tone is driven by former performance, whereas current tone has weak impact on future performance.

3.2 PCA Score

The sentiment score used in the previous portfolio formation depends solely on two ratios only, i.e. number of positive (negative) terms to the total number of words in the report. To gain a broader perspective, we deploy a principle component analysis (PCA) approach. In particular, we combine the 6 dictionary ratios into 1 one factor using PCA. This is conducted on a rolling window basis. Starting at the end of Dec 1988, we use the recent 60 months (included) to estimate the covariance matrix for of the dictionary ratios. Then, we extract the first eigenvector, W_t representing the first PCA component and map the 6 dictionary ratios into one using W_t. In the following month, we roll the 60 months window ahead, discarding the first the month in the data while incorporating recent info from Jan 1989. We repeat the process over the whole data, resulting in a time series of 349 months. Eventually, this results in W_t for $t = 1, ..., 349$ months. Similar to (2.1), the PCA sentiment score is given by

$$Z_{i,t}^{PCA} = X_{i,t}' W_t \tag{3.1}$$

where $X_{i,t}$ is a 6×1 column vector denoting the six dictionary ratios for a firm i at month t, while W_t is 6×1 column vector representing the first principle component estimated at month t using 60 months of data.

Table 4 provides summary statistics of the weights of the PCA to each dictionary as well as the proportion of variance explained by the first component. A number of comments are worth mentioning. First, we note that the first component explains, on average, 44% of the variability in the dictionary ratios. This proportion ranges between 35% and 55% at most. Second, similar to the Z score from (2.1), we note that, on average, positive sentiment takes a positive weight, whereas negative sentiment takes a negative weight. Third, the PCA weights for uncertainty and litigation also have negative weights, whereas valuation and alternative dictionaries exhibit positive coefficients. In terms of magnitude, we note that the positive and negative dictionary coefficients exhibit the largest magnitudes.[5]

Table 4: **Principle Component Analysis of Dictionaries**
This table reports the PCA results on a rolling window basis. In particular, it reports the first principle component, which we refer to as the vector of weights to map six dictionary ratios into a single index. In addition, the last row reports the total variation explained by the first principle component.

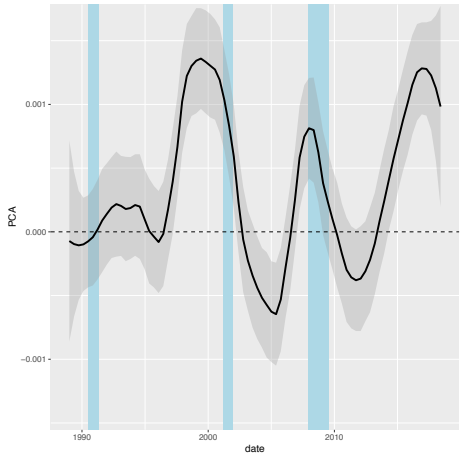
Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Positive	355	0.117	0.440	−0.512	−0.421	0.487	0.512
Negative	355	0.134	0.522	−0.574	−0.529	0.536	0.573
Uncertainty	355	0.079	0.382	−0.484	−0.412	0.428	0.482
Litigation	355	0.074	0.320	−0.476	−0.299	0.374	0.465
Valuation	355	0.075	0.338	−0.507	−0.260	0.370	0.506
Alternative	355	0.080	0.344	−0.473	−0.285	0.389	0.471
% Explained Variance	355	0.337	0.026	0.295	0.318	0.367	0.388

In Figure 4, we demonstrate the PCA score from (3.1) over time using a smoothed local regression. We observe that the score is cyclical over time. Similar to Figure 2 Panel (d), we see there is an overall increase until 2000. Afterwards, the score drops but increases again until we head to the 2007-09 financial crisis. After which, it appears the PCA score is on decline until a new upward trend begins in 2012.

As a robustness check, we repeat the same analysis from Table 3, where we group firms based on PCA scores into 4 portfolios and compute the next/current risk adjusted

[5]Note that by construction the second norm of the principle component sum to 1.

Figure 4: **PCA Dictionary Score**
The following figure demonstrates the dictionary score constructed using PCA analysis, which is the sum of six dictionary ratios, including positive, negative, uncertainty, litigation, valuation„ and alternative dictionaries. The figure is constructed using a local regression with a span parameter of 0.25. The lower and upper bounds illustrate the bottom and top 10% of the PCA score, respectively.



(a) PCA Index

return of each portfolio. In Table 5, we report the risk-adjusted returns alone in line with Table 3. Overall, we find consistent evidence with Tables 3 and **??**. Specifically, there is a monotonic increase in the portfolio current returns as the PCA sentiment score increases. The difference between the third and first tercile, however, while larger in economic magnitude, is less significant than the previous results.

3.3 Machine Learning Portfolio Formation

The former portfolio formation is done using a cutting point given the net tone, measured using the score from (2.1) or the PCA approach from (3.1) implemented in the former section. Nevertheless, the dictionary weighting in either approach is unsupervised, i.e. the construction of either sentiment score is conducted independently of market data. In this section, we provide a supervised approach, in which the dictionaries are weighted with respect to market data.

Table 5: **Portfolio Sorting - PCA Score**
This table presents estimates from OLS regression of monthly equally-weighted excess returns on each sentiment-sorted portfolio on the three Fama and French (1993) and Carhart (1997) momentum risk factors. Portfolios are sorted on a monthly basis with respect to the PCA score from (3.1). Alphas are reported on an annual basis, and standard errors are adjusted for heteroskedasticity and autocorrelation using Newey and West (1986) with three lags.

	(1)	(2)	(3)	(3)-(1)
	Panel (a) Value Weighted			
Next Month	−0.006	0.022**	0.017	0.023
	(0.015)	(0.009)	(0.015)	(0.022)
Current Month	0.064***	0.095***	0.076***	0.012
	(0.014)	(0.012)	(0.013)	(0.019)
	Panel (b) Equally Weighted			
Next Month	0.063***	0.043***	0.082***	0.019
	(0.013)	(0.011)	(0.011)	(0.013)
Current Month	0.053***	0.069***	0.069***	0.016
	(0.011)	(0.011)	(0.014)	(0.014)
Note:				*p<0.1; **p<0.05; ***p<0.01

3.3.1 Elastic Net Regression

Similar to Equation (3.1), we attempt to find the "optimal" weights to map the feature space $X_{i,t}$ into a single sentiment score. The feature space $X_{i,t}$ constitutes a textual summary of the analyst reports covering firm i at month t, which covers the relative frequency of each of the six dictionaries summarized in Figures 2 and 3. Unlike the PCA score from (3.1), the following score takes into account firms' abnormal returns.

At each month, we deploy a linear regression model with net elastic penalty to choose the optimal feature space. The dependent variable is the abnormal stock return estimated as the alpha (intercept) from the Fama and French (1993) and Carhart (1997) four risk factors pricing model using a rolling window of 60 months. The regressors (feature space) are given by the six dictionary ratios. The net elastic penalty penalizes the first and the second norm of the weights allocated to each dictionary. The first one serves as a elimination (selection) process, whereas the second norm serves as a shrinkage approach. An in-between such as the net elastic is expected to serve for selection as well as robust estimation of the dictionary weights. Formally, we set the net elastic parameter to 0.5, where 0 denotes a ridge regression (shrinking the second norm toward zero) and 1 denotes the Lasso regression (shrinking the first norm toward zero).

To implement this, we refer to the algorithm proposed by Friedman et al. (2010) and, hence, the `glmnet R` library to estimate the model. At each month, we fit the model in a cross-sectional manner with 10-folds cross validation (CV). This allows us to find the optimal parameter for the penalty constraint, which is denoted by λ. Given the CV, we find the optimal λ^{CV} that yields the lowest mean-squared error (MSE). In particular, the dictionary weights is the vector β that minimizes the following objective

$$\frac{1}{2N_t}\sum_{i=1}^{N_t}\left(\alpha_{i,t} - X'_{i,t-1}\beta\right)^2 + \lambda\left[(1-\theta)\;||\;\beta\;||_2^2\;/2 + \theta\;||\;\beta\;||_1\right] \tag{3.2}$$

with N_t is the number of stocks in month t, $\alpha_{i,t}$ is the abnormal return of stock i in month t estimated using a rolling window of 60 months, $X_{i,t-1}$ is the lagged feature space which is a 6×1 vector covering the six dictionary ratios, λ is the penalty parameter, and θ is the net elastic parameter, which we set $\theta = 0.5$.

Note from (3.2) if $\theta = 1$, then the penalty serves as a constraint on the first norm of β. This corresponds to the LASSO regression. On the other hand, if $\theta = 0$, then the penalty acts as a constraint on the second norm. The latter corresponds to the Ridge regression. Hence, a net elastic with $\theta = 0.5$ penalizes the estimated weight with respect to both constraints.

3.3.2 Solution and Score

We minimize (3.2) with respect to λ^{CV}, which is tuned using the 10-folds CV, for each month t in the data. Since we need 60 months to compute the abnormal returns, the sample starts from 1988-12-31. We denote the solution to the optimization problem from (3.2) at month t by $\hat{\beta}_t^{CV}$. As a result, the textual value of firm i at month t is given by

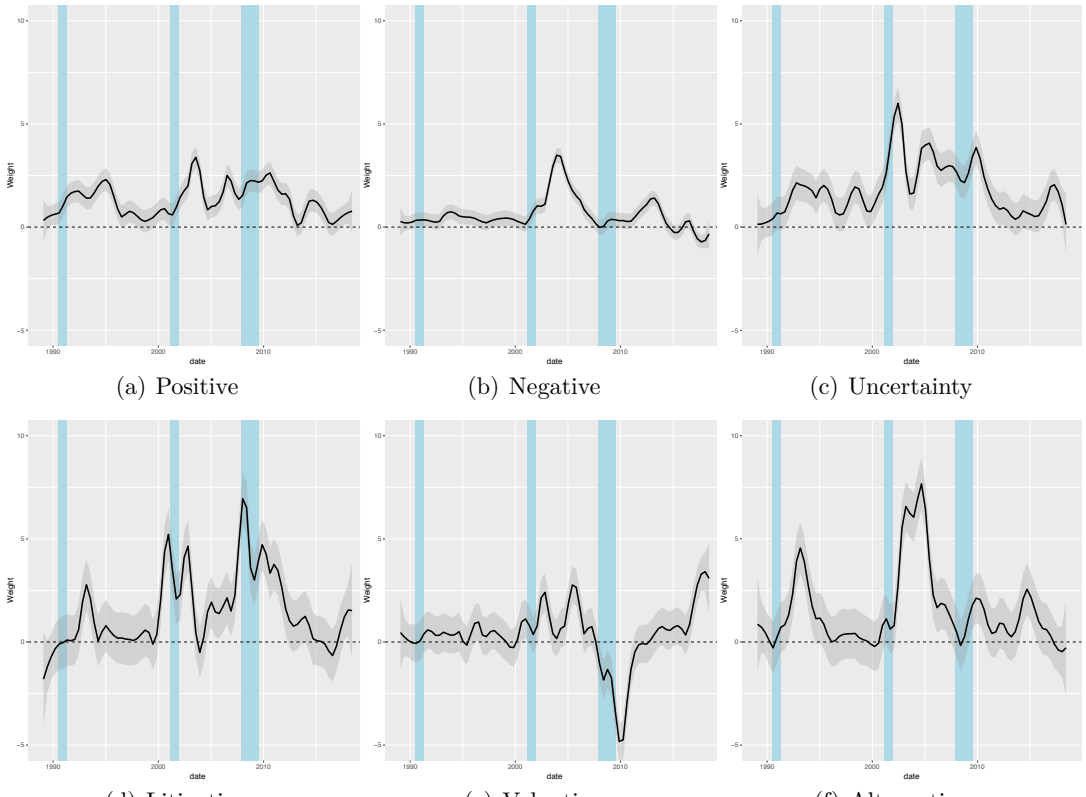
$$Z_{i,t}^{CV} = X'_{i,t}\hat{\beta}_t^{CV} \tag{3.3}$$

In Figure 5, we illustrate the weight attributed to each dictionary over time. For all dictionaries (all panels), we witness significant time series variation of the weights. Since the model is net elastic, dictionaries get either dropped over time or shrank toward zero,

as we observe from all panels. Nonetheless, the dictionaries with the largest magnitude are litigation and uncertainty. In most cases, there is a positive relationship between the dictionaries and abnormal returns. The only exception is for valuation terms, for which we witness negative magnitude before the early 2000s recession and later during/after the 2007-09 financial crisis.

Comparing between positive and negative sentiment in Panel (a) and (b) from Figure 5, respectively, we note that positive sentiment exhibits consistently larger impact in terms of magnitude. Additionally, while the weight of the negative dictionary increases after the early 2000s, it appears that negative sentiment has less importance in the post 2007-09 financial crisis period.

Figure 5: **Weights of Dictionaries over Time using Net Elastic Regression**:
The following panels denote the smoothed weight of each dictionary over time. The weight of each denotes the importance of each dictionary supervised using market data (stock returns). The bars denote recession periods identified according to the National Bureau of Economic Research (NBER). The horizontal dashed line is the zero level, where the solid line is the weight attributed to the corresponding dictionary. The latter is smoothed using a local regression with a span of 0.3.



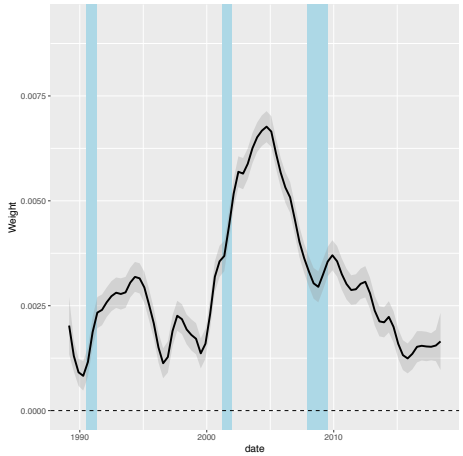
(a) Positive (b) Negative (c) Uncertainty

(d) Litigation (e) Valuation (f) Alternative

Putting the individual ML dictionary results together, we plot the average score from (3.3) in Figure 6. We observe that the score exhibits cyclicality. It appears to increase toward 2005 and exhibits a decline toward the end of the sample. Note that the ML score exhibits a pattern that is similar, albeit inversely, to the PCA. This is mainly due to the fact that the score is a linear combination of the six dictionaries. Since the combination is achieved from the net elastic regression using the abnormal returns, we expect the score to capture the textual content of the analyst report with respect to market reaction. For this reason, we refer to the score as supervised, unlike the case for the PCA in which the weighting of the dictionaries is independent of market reaction.

The approach taken in this section is data driven and in line with the one proposed by Jegadeesh and Wu (2013). Different from theirs, we deploy the weighting scheme on the dictionary level rather than the word level. Additionally, the authors study the market reaction to 10-K annual disclosures by public firms as opposed to reports on public firms generated by sell-side analysts. Their reaction is captured using abnormal returns as as the cumulative stock return over the value weighted market portfolio over four days around and including the day of the 10-K disclosure.

Figure 6: **ML Sentiment Score**
This figure demonstrates the average textual score from Equation (3.3) over time.



3.3.3 Portfolio Formation

Given the ML score on the firm-month level, we deploy the same portfolio test as before but using the textual score from Equation (3.3). In particular, at each month t, we group firms based on $Z_{i,t}^{CV}$ into terciles and construct both value and equally weighted portfolios for the next (current) period. In line with Table 5, we report the results in Table 6.

We observe that there is both economically and statistically significant evidence for the value-weighted portfolios from in Panel (a) of Table 6. Particularly, we find that stocks ranked in the higher tercile outperform those ranked in the lower tercile by 2.2% in annual basis. However, this improvement is statistically insignificant. Nonetheless, we also find no evidence that the analysts textual content is driven by the market as Panel (b) indicates. Unlike the naive net tone sentiment index, the ML index avoids overfitting and exhibits no evidence that sell-side analyst report content is driven by the market.

Table 6: **Portfolio Sorting - Machine Learning Score**
This table presents estimates from OLS regression of excess returns on each sentiment-sorted portfolio on the three Fama and French (1993) and Carhart (1997) momentum risk factors. Portfolios are sorted on a monthly basis with respect to the ML score from (3.3) and value and equally-weighted in Panel (a) and (b), respectively. Alphas are reported on an annual basis, and standard errors are adjusted for heteroskedasticity and autocorrelation using Newey and West (1986) with three lags.

	(1)	(2)	(3)	(3)-(1)
	Panel (a) Value Weighted			
Next Month	−0.013	0.034***	0.009	0.022
	(0.014)	(0.013)	(0.010)	(0.018)
Current Month	0.081***	0.074***	0.092***	0.011
	(0.016)	(0.013)	(0.015)	(0.019)
	Panel (b) Equally Weighted			
Next Month	0.077***	0.051***	0.049***	−0.028**
	(0.013)	(0.013)	(0.010)	(0.012)
Current Month	0.067***	0.053***	0.057***	−0.009
	(0.013)	(0.012)	(0.012)	(0.015)
Note:				*p<0.1; **p<0.05; ***p<0.01

When one considers the case for the equally-weighted portfolios, however, we observe a different results. In fact, we find that firms ranked in the higher tercile underperform those ranked in the lowest tercile by 2.8% per annum, which is significant at the 5%

level. This indicates that this out-performance is also size dependent. One argument is that large stocks receive more attention and, hence, get greater reaction. To address this we repeat the same analysis using independent double portfolio sorting using size, i.e. sort on size and then on ML score. The results are reported in Table A.1 in the appendix. For the large cap stocks, we observe a significant positive market reaction over the next month, when using equal portfolio weights. On the other hand, we observe observe an opposite market reaction to the analyst reports. Nevertheless, we also observe that analyst covering small cap stocks are also driven by the market, in which we find a strong evidence of contemporaneous returns in line with the findings from Panel (b) of Table 3.

The main takeaway from the above analysis is that the results and conclusions are sensitive to the construction of the textual score. In the naive sentiment case, we note that assigning ad-hoc weights to each dictionary can be viewed as an overfitting problem. On the other hand, a data-driven extracted score avoids this and provides different insights. The main challenge with the ML aspect, is that is unclear what the score constitutes.

3.3.4 What Stands Behind the ML Score?

ML tools can be deemed as a black box, even for the simplest algorithm implementations. This is mainly due to the selection process that picks the most important factors in the cross-validation manner. Nonetheless, as Figure 5 implies, we can still uncover the important factors that feed into the algorithm. To gain a statistical perspective on the constituents of the ML score, we run a panel regression with firm fixed effects using the proposed ML index. In particular, the dependent variable is the $Z_{i,t}^{CV}$, whereas the regressors are the six dictionaries. This is rather a reverse engineering exercise to understand the driving factors behind the proposed ML score. In addition, we run the same model for the net tone and PCA scores.

In Table 7, we report the estimated coefficients from the panel regression. Unsurprisingly, we note that the positive and negative coefficients in the net tone score correspond to 1 and -1, respectively, with an $R^2 = 1$. On the other hand, the results for the PCA

Table 7: **Dictionary Weighting Panel Regression**
This table reports the coefficients of a panel regression with firm fixed effects, in which the dependent variable is the dictionary score. Columns 1, 2, and 3 correspond to the net tone Z, the PCA score Z^{PCA}, and the ML score Z^{CV}, respectively. In each case, the regressors are given by the six dictionaries, which denote the mean relative frequency of specific terms appearing in the textual report. Similar to column 3, columns 4 and 5 conduct the same test, however, for sub-sample of small cap and large cap, respectively.

	Dependent variable:				
	Z	Z^{PCA}	Z^{CV}	Z^{CV} Small	Z^{CV} Large
	(1)	(2)	(3)	(4)	(5)
Positive	1.000***	0.339***	0.885***	0.974***	0.827***
	(0.000)	(0.004)	(0.009)	(0.011)	(0.013)
Negative	−1.000***	0.112***	0.651***	0.440***	0.790***
	(0.000)	(0.004)	(0.009)	(0.011)	(0.012)
Uncertainty	0.000	−0.100***	2.118***	2.243***	2.040***
	(0.000)	(0.007)	(0.015)	(0.021)	(0.020)
Litigation	−0.000	0.161***	0.864***	0.724***	0.980***
	(0.000)	(0.015)	(0.030)	(0.041)	(0.042)
Valuation	0.000	0.224***	−0.174***	0.085***	−0.262***
	(0.000)	(0.009)	(0.019)	(0.024)	(0.027)
Alternative	−0.000	0.335***	0.478***	0.592***	0.376***
	(0.000)	(0.008)	(0.018)	(0.024)	(0.026)
Firm FE	Yes	Yes	Yes	Yes	Yes
Observations	216,780	208,567	155,974	70,096	85,878
Adjusted R^2	1.000	0.069	0.280	0.365	0.238
Note:					*p<0.1; **p<0.05; ***p<0.01

and ML vary, since both are extracted in a data-driven manner on a rolling window basis. Since the ML score is extracted on the cross-sectional level at each month, we observe that the R^2 is substantially larger than that reported for the PCA. At the same time, since we deploy an elastic net penalty over time, coefficients also vary over time. However, the coefficients from the panel regression uncover, on average, the constituents of the ML score. For instance, we observe that Z^{CV} has positive loading on all dictionaries except that for valuation related words.

In columns 4 and 5 in Table 7, we split the sample into small and large cap in line with Panel (c) from Table 6. We notice a couple of key differences. First, we observe that the ML score has greater loading on the negative dictionary for large caps, as compared with small caps. Second, we note that the negative coefficient of the valuation dictionary is mainly evident for large caps but not for small caps. In other words, this indicates that greater use of valuation terms for larger caps is associated with lower future abnormal

returns, whereas the opposite holds true for small caps. Size-based differences across the remaining dictionaries are also present. Each of these pieces of evidence demonstrate the heterogeneity in textual content extraction in terms of what is relevant to explain abnormal returns.

4 Panel Analysis

To this point, the previous analysis is conducted using non-parametric tests (i.e. portfolio formation). To control for other factors and implications of the extracted textual scores, we conduct a battery of parametric tests using panel data regression. In all cases, we investigate the significance of the three dictionary scores on firm short interest and return volatility, after controlling for additional factors, such as firm characteristics and firm-quarter fixed effects.

4.1 Summary Statistics

After constructing the three textual indices on the firm-month level, we construct our final panel on the quarter-firm level to merge with other financial data. In Table 8, we provide summary statistics of the panel.

Since the textual variables are computed as ratios, the three scores represent fractions. On average, we observe that the three scores are around zero. However, we observe that the ML score exhibits larger variation. In terms of size, the firms in this sample have an average of $6.44 billion with an average of nearly 8 sell-side analysts providing coverage per firm.

4.2 Short Interest

The percentage of shares that investors short sell can be an indicator of the market's sentiment of whether to expect a fall in value of the shares. Existing literature also finds a negative significant relationship between short interest and stock returns (see e.g., Asquith and Meulbroek (1995); Desai et al. (2002)) Engelberg et al. (2012) show that short sellers successfully process negative information in their trading strategies. Hence,

Table 8: **Summary Statistics of Target Variables**
This table provides summary statistics of the variables used in the panel analysis. The panel is at the firm-quarter level. The variables Z, Z^{PCA}, and Z^{CV} correspond to the naive, PCA, and ML extracted textual scores, respectively. Total analysts is the number of analysts covering the firm in a given quarter. Size is the natural logarithm of the book value of total assets. Leverage is the ratio of long term debts plus debt in current liabilities divided by stockholders book equity. Log(SI) is the natural logarithm of short interest and R_t and σ_t are the quarterly stock return and return volatility, respectively. In all cases, variables are winsorized at the 1% and 99% levels cross-sectionally on a quarterly basis.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Z	72,251	−0.0004	0.001	−0.012	−0.001	0.000	0.009
Z^{PCA}	69,092	0.0003	0.002	−0.007	−0.001	0.002	0.008
Z^{CV}	51,525	0.003	0.003	−0.007	0.001	0.004	0.039
Total Analysts	51,922	7.663	6.453	1.000	3.000	11.000	33.500
Size	67,453	1.863	0.353	0.438	1.667	2.112	2.532
ROA	67,432	0.001	0.049	−0.493	0.001	0.020	0.207
Tobin's q	66,892	1.972	1.593	0.420	1.075	2.189	26.165
Leverage	62,729	1.173	2.657	0.000	0.080	1.071	52.836
log(SI)	57,213	13.446	2.494	2.382	12.264	15.227	18.054
R_t	72,251	0.039	0.232	−1.194	−0.075	0.144	2.257
σ_t	72,251	0.183	0.146	0.013	0.089	0.230	1.878
Volume	72,251	11.601	2.071	5.020	10.188	13.057	16.845
Market Cap	72,251	13.441	2.063	8.166	11.955	14.837	19.091

does the information within analyst reports also predict short selling? Thus our goal here is also to investigate the predictive power of analyst reports on short-interest.

In Table 9, we investigate the textual content of analyst reports using each extracted index and examine whether or not each index has any value in explaining future short interest. Using a panel regression with firm-quarter fixed effects, we first regress the same quarter short interest on each index along with other controls to investigate the contemporaneous relationship (see columns 1,2, and 3). Second, we regress the next quarter short interest on the lagged explanatory variables to investigate the lead-lag relationship.

Starting with the naive sentiment score, we note that the coefficient on net tone is negative and significant at the 1% level for the contemporaneous and the next quarter. In either case, this evidence indicates that more positive tone relative to negative is associated with lower short selling in the current and next quarters. This is consistent with existing literature that indicates short selling is associated with negative news. For the PCA score, we find no statistical significance at all. For the ML score, we find a positive significant relationship between the score and current/next quarter short interest.

Table 9: **Panel Regression - Short Interest**
This table reports panel regressions with firm and quarter fixed effects. The dependent variable is the natural logarithm of the short interest, which is denoted by log(SI). In the first three columns, the results correspond to the same quarter short interest, whereas the last three columns correspond to the next quarter relative change. In all cases, significance levels are computed with respect to robust standard errors, which are reported in parentheses.

	Current log(SI)			*Next Quarter log(SI)*		
	(1)	(2)	(3)	(4)	(5)	(6)
Z	-34.450^{***}			-39.642^{***}		
	(5.162)			(4.955)		
Z^{PCA}		-1.366			7.543	
		(5.848)			(5.713)	
Z^{CV}			16.926^{***}			16.369^{***}
			(2.048)			(1.966)
Total Analysts	0.026^{***}	0.028^{***}	0.021^{***}	0.023^{***}	0.024^{***}	0.019^{***}
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
ROA	-1.517^{***}	-1.535^{***}	-1.345^{***}	-1.345^{***}	-1.334^{***}	-1.056^{***}
	(0.175)	(0.175)	(0.186)	(0.172)	(0.172)	(0.182)
Tobin's q	0.159^{***}	0.160^{***}	0.154^{***}	0.188^{***}	0.187^{***}	0.171^{***}
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Leverage	0.015^{***}	0.016^{***}	0.014^{***}	0.015^{***}	0.016^{***}	0.015^{***}
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Return	-0.260^{***}	-0.267^{***}	-0.254^{***}	-0.206^{***}	-0.210^{***}	-0.215^{***}
	(0.028)	(0.028)	(0.029)	(0.028)	(0.028)	(0.029)
Size	4.472^{***}	4.450^{***}	3.823^{***}	4.280^{***}	4.241^{***}	3.679^{***}
	(0.097)	(0.098)	(0.103)	(0.093)	(0.094)	(0.100)
Observations	39,930	38,925	32,045	39,773	38,998	32,072
Adjusted R^2	0.108	0.107	0.073	0.105	0.103	0.070
Note:						*p<0.1; **p<0.05; ***p<0.01

To better understand the ML result, recall that the ML score is extracted with respect to future abnormal returns. Mainly, for large (small) cap stocks the relationship is positive (negative), such that a higher ML score is associated with higher (lower) future abnormal returns. Moreover, Rapach et al. (2016) show that short interest is a strong predictor of stock returns and that a one-standard-deviation increase in the aggregate short interest level corresponds to a 6% decrease in the future annualized market excess return. Hence, given the findings from portfolio formation and evidence documented by Rapach et al. (2016), we expect that the ML score should have a negative (positive) loading on current/future short interest for large (small) stocks. In Tables A.2 and A.3, we repeat the same panel regression but with respect to small and large stocks, respectively. We find that the relationship is positive and strongly statistically significant only for the

small cap stocks with an adjusted R^2 of around 16%. This is in contrast to large cap stocks where panel regressions with ML scores explain less than 1% of variation in short interest.

4.3 Stock Volatility

In Table 10, we report panel analysis for realized stock return volatility. Our specification is similar to Table 9 though we add additional controls for lagged volatility. The results are consistent across current quarter (see columns 1,2, and 3) and next quarter (columns 4,5, and 6) volatility: ML scores significantly explain volatility but simple sentiment and PCA scores do not. We further break down our analysis by market cap in Tables A.4 and A.5. We find weak evidence that ML scores predict next quarter volatility for small cap firms but no relationship for either contemporaneous or subsequent volatility in large cap firms.

5 Conclusion

We document substantial time series variation in analyst report characteristics and content across 4 existing Loughran and McDonald (2011) and two new dictionaries related to valuation and alternative data. We employ standard sentiment, principle components analysis (PCA), and machine learning (ML) methods in rolling windows across the time series to predict equity returns, short interest, and volatility. Our findings, overall, indicate that there is a different market reaction to analyst reports for small and large cap stocks. ML based results in particular suggest that the information within sell-side analyst reports is particularly valuable to the market participants for small firms in which the information environment is not as well developed. However, declining trading commissions have reduced incentives for sell-side analysts to cover small firms. The U.S. Securities and Exchange Commission has tried to stem this tide though tick size pilot programs that have been mostly unsuccessful.[6] However, our ML results suggest that

[6]See "Congress' Failed Stock Market Experiment Cost Investors $900 Million" at Barron's via https://www.barrons.com/articles/sec-tick-size-pilot-program-1536961160

Table 10: **Panel Regression - Volatility**
This table reports the panel regression with firm and quarter fixed effects. The dependent variable is the stock quarterly realized volatility. Quarterly realized volatility is calculated as the square root of the sum squares of the corresponding thee months return. In the first three columns, the results correspond to the current quarter volatility, whereas the last three columns correspond to the next quarter volatility. In all cases, significance levels are computed with respect to robust standard errors, which are reported in parenthesis.

	Current Quarter Volatility			*Next Quarter Volatility*		
	(1)	(2)	(3)	(4)	(5)	(6)
Z	−0.616			0.157		
	(0.503)			(0.541)		
Z^{PCA}		0.377			0.833	
		(0.522)			(0.587)	
Z^{CV}			−0.399**			−0.736***
			(0.190)			(0.207)
Total Analysts	−0.003***	−0.003***	−0.003***	−0.002***	−0.002***	−0.002***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
ROA	−0.303***	−0.305***	−0.268***	−0.202***	−0.201***	−0.115***
	(0.024)	(0.024)	(0.027)	(0.026)	(0.026)	(0.029)
Tobin's q	−0.002**	−0.002*	−0.005***	0.003***	0.003***	−0.0001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Leverage	0.002***	0.002***	0.002***	0.001***	0.001***	0.002***
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0004)
Return	0.167***	0.167***	0.159***	−0.060***	−0.061***	−0.060***
	(0.006)	(0.006)	(0.006)	(0.004)	(0.004)	(0.005)
Size	−0.146***	−0.144***	−0.164***	−0.099***	−0.099***	−0.108***
	(0.008)	(0.009)	(0.010)	(0.009)	(0.009)	(0.011)
Lagged Volatility	0.123***	0.123***	0.103***	0.214***	0.215***	0.205***
	(0.007)	(0.007)	(0.008)	(0.009)	(0.009)	(0.010)
Volume	0.044***	0.045***	0.048***	0.010***	0.011***	0.011***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	45,962	44,600	35,210	46,004	44,850	35,104
Adjusted R^2	0.191	0.192	0.174	0.051	0.052	0.035
Note:						*p<0.1; **p<0.05; ***p<0.01

regulators should continue to improve market structures to encourage sell-side analyst information production particularly for small public firms.

In addition, the ML findings also stress the importance of using further advanced tools to extract textual information. Our results suggest that supervised machine learning consistently extracts information used to explain important market reactions to analyst reports, particularly in contrast to PCA methods. On one hand, naive net tone approach is transparent and easier to comprehend. On the other hand, it is critical to find optimal words for different terms to better capture the semantics of a textual corpus. While our

analysis is limited to dictionaries use only, the data-driven approach can be extended to analyze specific words and deploy more advanced tools such as word2vec (see e.g., Mikolov et al. (2013)). We leave both for future research.

References

Asquith, P., Meulbroek, L. K., 1995. An empirical investigation of short interest. Division of Research, Harvard Business School.

Asquith, P., Mikhail, M. B., Au, A. S., 2005. Information content of equity analyst reports. Journal of Financial Economics 75, 245 – 282.

Brennan, M. J., Hughes, P. J., 1991. Stock prices and the supply of information. The Journal of Finance 46, 1665–1691.

Brown, L. D., Call, A. C., Clement, M. B., Sharp, N. Y., 2015. Inside the "black box" of sell-side financial analysts. Journal of Accounting Research 53, 1–47.

Carhart, M. M., 1997. On persistence in mutual fund performance. The Journal of Finance 52, 57–82.

Chen, X., Cheng, Q., Lo, K., 2010. On the relationship between analyst reports and corporate disclosures: Exploring the roles of information discovery and interpretation. Journal of Accounting and Economics 49, 206–226.

De Franco, G., Hope, O.-K., Vyas, D., Zhou, Y., 2015. Analyst report readability. Contemporary Accounting Research 32, 76–104.

Desai, H., Ramesh, K., Thiagarajan, S. R., Balachandran, B. V., 2002. An investigation of the informational role of short interest in the nasdaq market. The Journal of Finance 57, 2263–2287.

Engelberg, J. E., Reed, A. V., Ringgenberg, M. C., 2012. How are shorts informed?: Short sellers, news, and information processing. Journal of Financial Economics 105, 260–278.

Fama, E. F., French, K. R., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33, 3–56.

Frankel, R., Kothari, S., Weber, J., 2006. Determinants of the informativeness of analyst research. Journal of Accounting and Economics 41, 29–54.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33, 1–22.

Hong, H., Lim, T., Stein, J. C., 2000. Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies. The Journal of Finance 55, 265–295.

Huang, A. H., Zang, A. Y., Zheng, R., 2014. Evidence on the information content of text in analyst reports. The Accounting Review 89, 2151–2180.

Jegadeesh, N., Wu, D., 2013. Word power: A new approach for content analysis. Journal of Financial Economics 110, 712–729.

Lehavy, R., Li, F., Merkley, K., 2011. The effect of annual report readability on analyst following and the properties of their earnings forecasts. The Accounting Review 86, 1087–1115.

Loughran, T., McDonald, B., 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. The Journal of Finance 66, 35–65.

Merkley, K., Michaely, R., Pacelli, J., 2017. Does the scope of the sell-side analyst industry matter? an examination of bias, accuracy, and information content of analyst reports. The Journal of Finance 72, 1285–1334.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 .

Newey, W. K., West, K. D., 1986. A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix.

Painter, M., 2018. Unlevelling the playing field: The investment value and capital market consequences of alternative data. Available at SSRN 3222741 .

Previts, G. J., Bricker, R. J., Robinson, T. R., Young, S. J., 1994. A content analysis of sell-side financial analyst company reports. Accounting Horizons 8, 55.

Rapach, D. E., Ringgenberg, M. C., Zhou, G., 2016. Short interest and aggregate stock returns. Journal of Financial Economics 121, 46–65.

A Additional Results

Table A.1: **Portfolio Sorting - Machine Learning Score**
This table presents estimates from OLS regression of excess returns on each sentiment-sorted portfolio on the three Fama and French (1993) and Carhart (1997) momentum risk factors. Portfolios are sorted on a monthly basis with respect to the ML score from (3.3) and value and equally-weighted in Panel (a) and (b), respectively. Alphas are reported on an annual basis, and standard errors are adjusted for heteroskedasticity and autocorrelation using Newey and West (1986) with three lags.

	(1)	(2)	(3)	(3)-(1)
	Panel (a) Value Weighted - Large Cap Stocks			
Next Month	−0.015	0.035***	0.015	0.030
	(0.013)	(0.011)	(0.012)	(0.019)
Current Month	0.074***	0.069***	0.098***	0.024
	(0.016)	(0.012)	(0.018)	(0.023)
	Panel (b) Equally Weighted - Large Cap Stocks			
Next Month	0.003	0.036***	0.040***	0.037***
	(0.012)	(0.013)	(0.011)	(0.012)
Current Month	0.064***	0.073***	0.079***	0.015
	(0.014)	(0.011)	(0.013)	(0.013)
	Panel (c) Value Weighted - Small Cap Stocks			
Next Month	0.069***	0.031*	0.061***	−0.007
	(0.018)	(0.016)	(0.018)	(0.022)
Current Month	0.111***	0.060***	0.037*	−0.074***
	(0.021)	(0.019)	(0.019)	(0.027)
	Panel (d) Equally Weighted - Small Cap Stocks			
Next Month	0.132***	0.089***	0.086***	−0.046*
	(0.020)	(0.018)	(0.019)	(0.024)
Current Month	0.078***	0.037*	0.005	−0.073***
	(0.020)	(0.020)	(0.019)	(0.026)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table A.2: **Panel Regression - Short Interest for Small Cap**
This table reports the panel regression with firm and quarter fixed effects. The dependent variable is the natural log of short interest, which denoted by log(SI). In the first three columns, the results correspond to the same quarter log(SI), whereas the last three columns correspond to the next quarter log(SI). In all cases, significance levels are computed with respect to robust standard errors, which are reported in parenthesis.

	Current log(SI)			*Next Quarter log(SI)*		
	(1)	(2)	(3)	(4)	(5)	(6)
Z	−8.630			−25.652***		
	(10.034)			(9.526)		
Z^{PCA}		2.468			14.023	
		(10.139)			(9.773)	
Z^{CV}			15.813***			17.729***
			(4.109)			(3.889)
Total Analysts	0.143***	0.144***	0.138***	0.132***	0.134***	0.134***
	(0.005)	(0.005)	(0.006)	(0.005)	(0.005)	(0.006)
ROA	−0.833***	−0.795***	−0.836***	−0.760***	−0.712***	−0.651***
	(0.223)	(0.224)	(0.246)	(0.219)	(0.219)	(0.239)
Tobin's q	0.298***	0.299***	0.296***	0.345***	0.344***	0.328***
	(0.013)	(0.013)	(0.014)	(0.013)	(0.013)	(0.014)
Leverage	0.011***	0.010**	0.008*	0.011***	0.011***	0.008*
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Return	−0.362***	−0.371***	−0.325***	−0.184***	−0.186***	−0.148***
	(0.041)	(0.042)	(0.044)	(0.042)	(0.042)	(0.044)
Size	4.623***	4.592***	4.443***	4.511***	4.466***	4.341***
	(0.153)	(0.153)	(0.168)	(0.146)	(0.146)	(0.162)
Observations	15,424	15,128	11,487	15,405	15,170	11,504
Adjusted R^2	0.165	0.168	0.154	0.167	0.168	0.156
Note:						*p<0.1; **p<0.05; ***p<0.01

Table A.3: **Panel Regression - Short Interest for Large Cap**
This table reports the panel regression with firm and quarter fixed effects. The dependent variable is the natural log of short interest, which denoted by log(SI). In the first three columns, the results correspond to the same quarter log(SI), whereas the last three columns correspond to the next quarter log(SI). In all cases, significance levels are computed with respect to robust standard errors, which are reported in parenthesis.

	Current log(SI)			*Next Quarter log(SI)*		
	(1)	(2)	(3)	(4)	(5)	(6)
Z	-23.535^{***}			-24.191^{***}		
	(5.232)			(5.097)		
Z^{PCA}		6.258			4.819	
		(5.980)			(6.032)	
Z^{CV}			4.188^{**}			3.741^{*}
			(1.951)			(1.932)
Total Analysts	0.027^{***}	0.029^{***}	0.020^{***}	0.024^{***}	0.026^{***}	0.018^{***}
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
ROA	-1.836^{***}	-1.907^{***}	-2.098^{***}	-1.689^{***}	-1.666^{***}	-1.902^{***}
	(0.211)	(0.211)	(0.238)	(0.208)	(0.207)	(0.236)
Tobin's q	0.033^{***}	0.033^{***}	0.023^{***}	0.049^{***}	0.046^{***}	0.033^{***}
	(0.006)	(0.006)	(0.007)	(0.006)	(0.006)	(0.007)
Leverage	0.018^{***}	0.018^{***}	0.022^{***}	0.018^{***}	0.018^{***}	0.022^{***}
	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.003)
Return	-0.098^{***}	-0.104^{***}	-0.101^{***}	-0.236^{***}	-0.244^{***}	-0.242^{***}
	(0.029)	(0.030)	(0.032)	(0.030)	(0.030)	(0.031)
Size	2.708^{***}	2.578^{***}	1.933^{***}	2.462^{***}	2.295^{***}	1.770^{***}
	(0.119)	(0.121)	(0.125)	(0.117)	(0.117)	(0.123)
Observations	24,506	23,797	20,558	24,368	23,828	20,568
Adjusted R^2	0.038	0.036	0.008	0.031	0.029	0.006
Note:						*p<0.1; **p<0.05; ***p<0.01

Table A.4: **Panel Regression - Volatility for Small Cap**
This table reports the panel regression with firm and quarter fixed effects. The dependent variable is the stock quarterly realized volatility. In the first three columns, the results correspond to the current quarter volatility, whereas the last three columns correspond to the next quarter volatility. In all cases, significance levels are computed with respect to robust standard errors, which are reported in parenthesis.

	Current Quarter Volatility			*Next Quarter Volatility*		
	(1)	(2)	(3)	(4)	(5)	(6)
Z	-1.784^{*}			0.668		
	(0.975)			(1.085)		
Z^{PCA}		0.795			1.885^{*}	
		(0.940)			(1.109)	
Z^{CV}			-0.318			-0.803^{*}
			(0.386)			(0.448)
Total Analysts	-0.005^{***}	-0.006^{***}	-0.006^{***}	-0.002^{***}	-0.002^{***}	-0.002^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
ROA	-0.255^{***}	-0.259^{***}	-0.259^{***}	-0.160^{***}	-0.158^{***}	-0.093^{**}
	(0.030)	(0.030)	(0.036)	(0.033)	(0.033)	(0.040)
Tobin's q	-0.010^{***}	-0.010^{***}	-0.011^{***}	-0.001	-0.001	-0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Leverage	0.002^{***}	0.002^{***}	0.002^{***}	0.002^{***}	0.002^{***}	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Return	0.175^{***}	0.175^{***}	0.178^{***}	-0.047^{***}	-0.048^{***}	-0.048^{***}
	(0.007)	(0.007)	(0.008)	(0.006)	(0.006)	(0.007)
Size	-0.139^{***}	-0.136^{***}	-0.141^{***}	-0.081^{***}	-0.081^{***}	-0.087^{***}
	(0.014)	(0.014)	(0.017)	(0.015)	(0.016)	(0.020)
Lagged Volatility	0.072^{***}	0.072^{***}	0.047^{***}	0.165^{***}	0.167^{***}	0.153^{***}
	(0.009)	(0.009)	(0.010)	(0.011)	(0.012)	(0.014)
Volume	0.055^{***}	0.056^{***}	0.059^{***}	0.011^{***}	0.011^{***}	0.010^{***}
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Observations	19,148	18,536	13,130	19,270	18,701	13,068
Adjusted R^2	0.183	0.184	0.178	0.001	0.001	-0.019
Note:						*p<0.1; **p<0.05; ***p<0.01

Table A.5: **Panel Regression - Volatility for Large Cap**
This table reports the panel regression with firm and quarter fixed effects. The dependent variable is the stock quarterly realized volatility. In the first three columns, the results correspond to the current quarter volatility, whereas the last three columns correspond to the next quarter volatility. In all cases, significance levels are computed with respect to robust standard errors, which are reported in parenthesis.

	Current Quarter Volatility			*Next Quarter Volatility*		
	(1)	(2)	(3)	(4)	(5)	(6)
Z	−0.366			−0.099		
	(0.532)			(0.546)		
Z^{PCA}		−0.184			0.005	
		(0.596)			(0.619)	
Z^{CV}			0.183			−0.277
			(0.206)			(0.213)
Total Analysts	−0.002***	−0.002***	−0.002***	−0.002***	−0.002***	−0.001***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
ROA	−0.335***	−0.339***	−0.241***	−0.237***	−0.236***	−0.125***
	(0.034)	(0.035)	(0.036)	(0.040)	(0.040)	(0.042)
Tobin's q	0.002	0.002*	−0.002*	0.005***	0.005***	0.002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Leverage	0.001***	0.001***	0.001***	0.001***	0.001***	0.001***
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0004)
Return	0.164***	0.163***	0.146***	−0.061***	−0.061***	−0.056***
	(0.009)	(0.009)	(0.009)	(0.006)	(0.006)	(0.006)
Size	−0.181***	−0.179***	−0.194***	−0.139***	−0.139***	−0.122***
	(0.013)	(0.014)	(0.014)	(0.015)	(0.015)	(0.016)
Lagged Volatility	0.111***	0.111***	0.101***	0.200***	0.201***	0.194***
	(0.011)	(0.011)	(0.011)	(0.014)	(0.014)	(0.015)
Volume	0.047***	0.047***	0.050***	0.016***	0.016***	0.018***
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
Observations	26,814	26,064	22,080	26,734	26,149	22,036
Adjusted R^2	0.193	0.193	0.166	0.060	0.060	0.044
Note:						*p<0.1; **p<0.05; ***p<0.01