Measuring Firms Investment Plans:

A text-based analysis

Abstract

By using an approach based on text data and supervised machine learning, this paper propose a novel measure of investment plans in the firm-level. I combine the procedure of Han, He, Rapach, and Zhou (2020) with the idea of flexible dictionary of Lima, Godeiro, and Mohsin (2020) and, I test a novel measure of investment plans based on text data from Management Discussion and Analysis (MD&A) disclosure in 10-K filings. In this study, the sample includes all US publicly traded firms in the period between January 1995 and December 2019. I build a unique dataset by merging information from multiple data sources. The annual firm-level financial and accounting data, I obtain from Compustat. The firms' 10-K filings are from the SEC Edgar database and the monthly US stock returns from the Center for Research in Security Prices (CRSP). The main find is that the words of MD&A matter to predict firm fundamentals, even in the case of investment growth, which is empirically challenging to measure in the firm-level.

Keywords: investment plans; expected investment growth; cross-section forecast; supervised machine learning

JEL Classification: G12, G14, G15

1 Introduction

Despite the literature provides support for the importance of investment plans in both aggregate-level and firm-level (Cochrane, 1991; Hou, Mo, Xue, & Zhang, 2020; Li, Wang, & Yu, 2020), this last one receives less attention since it is empirically challenging to find a reliable measure due to the investment plans are not observable (Lin & Lin, 2018). To address this issue, I explore relevant information from MD&A (Management Discussion and Analysis) disclosure in 10-K filings to propose a novel measure based on text data, which I do by combining the procedure of Han et al. (2020) with the idea of time varying dictionary developed by Lima et al. (2020).

Most of the studies choose to examine investment plans in aggregate level data (Lamont, 2000; Li et al., 2020) and empirical investigations of firm-level investment plans are exceptions. One of then use micro data available in a quarterly survey of Chief Financial Officers (CFOs) and consider expectations about future investment growth as a measure of investment plans (Gennaioli, Ma, & Shleifer, 2016). Recently, Hou et al. (2020) and Li and Wang (2018) try to predict future investment change by using the Fama and MacBeth (1973) cross-sectional predictive regression based on current variables.

Both approaches at the firm-level have limitations. For example, the data used by Gennaioli et al. (2016) is only available from 1998, which is a limitation for asset pricing studies (Li & Wang, 2018). The measure of Hou et al. (2020) may contain misspecification errors, since they use Fama and MacBeth (1973) approach (Lin & Lin, 2018). For Lin and Lin (2018), the expected investment change measure of Hou et al. (2020) is a poor proxy for future investment growth because of the limitations on the regression model and the potential choose of weak predictors.

To better illustrate this argument, Hou, Xue, and Zhang (2018) did an extensive empirical analysis of how the major pricing models explain the already documented anomalies. Their model, called q-factor, capture most of the anomalies. However, they find some that remain unexplained, including 46 not captured by the "q-factor" model (the q-anomalies). In an earlier version (NBER working paper 23394, May 2017), the authors suggest that q-anomalies may not be explained by the q-factor model because it ignores the inclusion of an expected

growth factor (an EIG factor related to investment plans), and also mention, that Hou, Xue, and Zhang (2015) option of not to include the EIG factor was due to concerns about the lack of reliable proxies for this variable.

In finance, the difficulty of measuring certain variables has been overcome with the use of more advanced techniques such as text regression and machine learning. For instance, Frankel, Jennings, and Lee (2016) analyzed Management's Discussion and Analysis (MD&A) in the 10-K report to identify the most important words to explain firm-level accruals. Manela and Moreira (2017) created an implied volatility measure based on news that made it possible to understand the relationship between risk and disasters concerns using a much larger sample than that provided by options implied volatility (VIX). Lima et al. (2020) analyzed textual data found on FED minutes and created a time-varying dictionary to predict GDP growth that allowed us to understand how predictive words change over time.

In this work, I propose a way to minimize the measurement problem of investment growth expectations by using the same cross-section forecast method of Han et al. (2020), but here with text data as did by Lima et al. (2020) in time-series domain. Hence, to suggest a novel measure for investment plans may be a substantial contribution. To be more specific, by using machine learning and text regression to predict investment plans, I add to a growing literature that applies machine learning tools to analyze economic questions (Mullainathan & Spiess, 2017). In addition, I also contribute to a better understanding of an asset pricing puzzle related to the role of expected investment growth in explain returns. If the EIG is indeed an important and new dimension of expected return (Hou et al., 2020), finding better ways to measure it is essential for the asset pricing literature.

2 Methodology

2.1 Sample and Data

I analyze the US publicly traded firms in the years between 1995 to 2019, due to the availability of 10-k filings. I build a unique dataset by merging information from multiple data sources. The annual firm-level financial and accounting data, I obtain from Compustat.

The firms' 10-K filings are from the SEC Edgar database. To analyze if investment plans are priced, I obtain monthly US stock returns from the Center for Research in Security Prices (CRSP).

In many cases, the MD&A section is incorporated by reference to the annual report, which is difficult to accurately parse since it usually appears in an exhibit that is part of the filing, the beginning and especially the ending position for the MD&A session typically is not obvious. So, as Loughran and McDonald (2011), I require at least 250 words in the MD&A section to leave the document at the sample.

2.2 Measure of Investment Plans

Since the firm's investment plans are not observable, I need to estimate it by using models that consider in each time t only the publicly available information up to time t. This is important because the investment theory assumptions predicts that the market investor use only available information in order to be able to price investment plans. In this sense, I estimate the benchmark measure using the classic Fama and MacBeth (1973) approach as the recent study of Hou et al. (2020), which I explain in sub section 2.3. For the text-based measures, I perform two main tests, the first one estimated each year (sub section 2.4) and the last one estimated by month (sub section 2.5). In the annual tests, I use four different approach to estimate the model, each one is explained in sub section 2.4. I also test the flexibility of the approach with others firms fundamentals (see sub section 3.1.3). However do to the limitation of the statistical tests in annual estimation, I propose to forecast the investment growth monthly, which I explain the procedure in sub section 2.5.

2.3 Benchmark Measure

The Equation (1) is the benchmarking of expected investment growth computed as Hou et al. (2020), referred to as the Expected Investment Growth of HMXZ (EIG_{HMXZ}). As shown in the Equation 1. Hou et al. (2020) used as predictors the log of Tobin's Q, a measure of operating cash flow, and the change in return on equity (dROE). This last one is an attempt to capture the short-term dynamic of the investment-to-assets change. Then the out-of-sample

prediction of change. The investment-to-assets changes are be obtained from the average slopes estimated from the prior 120-month rolling window with the most recent winsorized predictors. Here, I require a minimum of 30 months to estimate the EIG.

$$E_{i,t}[IG] = b_{0,t} + b_{dROE,t}dROE_{i,t-1} + b_{q,t}Q_{i,t-1} + b_{CF,t}CF_{i,t-1} + \epsilon_{i,t} \tag{1}$$

where:

$E_{i,t}[IG]$ = the change in investment-to-assets $(I/A_{i,t})$ year ending in calendar year t $(IG_{i,t} = I/A_{i,t} - I/A_{i,t-1})$;

$dROE_{i,t-1}$ = change in return on equity over the past four quarters;

$Q_{i,t-1}$ = the log of the market value of the firm divided by total assets in the fiscal year ending in calendar year $t-1$;

$CF_{i,t-1}$ = the operating cash flow in the fiscal year ending in calendar year $t-1$ divided by lag total assets.

2.4 Text-based Measure - Annual Estimation

To construct a text-based measure of investment plan, I use the model similar to Lima et al. (2020). They propose a method to enable the content of dictionaries to vary over time, making it entirely determined by the predictive power of its words, which maximizes the predictive ability of the dictionary, then is suitable to the problem of forecasting. One of advantages of this methodology, is that is not necessary a pres-specified fixed dictionary because the model decide from the data which words are more important to predict investment growth over time.

In order to do this procedure, I follow three steps: First, the words are transformed into numerical values, which create a high dimensional and sparse matrix. Second I use a supervised machine learning to reduct the dimensionality by selecting the most predictive words and use them to construct new predictor(s). Lastly, in the third step, the out-of-sample forecasts are made from the new predictor(s) selected in the prior step. This three procedure

is repeated recursively up to the end of the sample. In short, the content of the dictionary (the most predictive words) changes over time (Lima et al., 2020).

Step 1 - Pre-process the textual data

The first step in extracting meaningful information from textual data contained in the 10-k report, is to pre-process the text. In this work, the goal is to reduce the form of unstructured data to a numerical data readable by a statistical tool. In order to do this, first I grab all available 10-k reports from 1994 to 2019 in a collection of text, which the literature calls "corpus". Than I remove all stop words (e.g. also, but, did, would, etc), punctuation and numbers. After that, I perform a common natural language approach called stemming, which assigning morphological variants to common root words, for example, the words economic, economics, economically are all replaced by the common root economic.

After the pre-processing steps, I identify what the literature calls as collocations or n-grams. In this work I choose to identify collocations with no more then 2 words and whose frequency is above 100, this approach is similar to others research (Frankel et al., 2016; Lima et al., 2020; Manela & Moreira, 2017). Then I generate a vector X_s where each element shows the frequency that a given 1-word or 2-words phrases j appears on texts published at year t by firm i.

Thus, with no using a pre-determined word list (fixed dictionary) this step converts words into numerical values for each firm i and year t, although p is very large and some words are not observed for some individuals/periods. So, this numerical representation is high dimensional and sparse, which is not suitable to the classic approach used in previous work (George, Hwang, & Li, 2018; Hou, Mo, Xue, & Zhang, 2018; Li & Wang, 2018) and, dimension reduction techniques as (e.g., regularization, principal components analysis) can be a suitable solution.

However, before the dimension reduction I apply a tf-idf weight for each term as (Loughran & McDonald, 2011). The tf-idf is commonly used as a filter removes less important words either because they are rare or because they are too frequent (Gentzkow, Kelly, & Taddy, 2019). However (Loughran & McDonald, 2011) use as weighting scheme, which is useful to

consider all words and instead remove rare or too frequent word, I set a low value for that word. In addition, use tf-idf as a filter give to the researcher an ad-hoc cut-off to choose, which I avoid in this research and leave the data choose which words is important despite be rare or too frequent.

To compute each term weight consider N as the total number of 10-Ks in the sample, $tf_{i,j}$ as the raw count of the i^{th} word in the j^{th} document, df_i the number of documents containing at least one occurrence of the $i\text{-}th$ word, and a_j the average word count in the document j, then the weighted measure is:

$$w_{i,j} = \begin{cases} \frac{(1+log(tf_{i,j}))}{1+log(a_j)} log\frac{N}{df_i} & if \quad tf_{i,j} \geq 1 \\ 0 & othersise \end{cases} \tag{2}$$

The first term attenuates the influence of high-frequency words with a log transformation. For instance, the word "adverse" appears 28776 times in the sample while the word phrase "credit loss" appears only 20 times. It is unlikely that the influence of the collocation "adverse" is more than 1438 times that of "credit loss". The second term of equation (2) alters the impact of a word phrase based on its commonality. For example, the word "adverse" appears in more than 80% of the documents, which implies that the second term of equation will decrease the first term by more than 80%. On the other hand, because "credit loss" appears in comparatively few documents, the second term of equation now rises the first by a factor of approximately six.

Step 2 - Select high predictive words

For the yearly estimation, I adapted the Lima et al. (2020) time-series approach and use a cross-section and industry estimation as Frankel et al. (2016), but instead of Support Vector Regression I use Elastic Net, which is simpler and suitable to our problem. To estimate a model to forecast the expected investment growth ($y_{i,t+h}$) in time h for each firm i, each year t I use a set of information available up to year t and estimate the equation (6).

$$y_{i,t|t-s} = \phi_t X'_{i,t-s} + \beta_{i,t} Z'_{i,t-s} + \varepsilon_{i,t} \tag{3}$$

Where $X_{i,t}$ is the $p \times 1$ vector with p traditional variables, $Z_{i,t}$ is the $k \times 1$ vector with k word count for each i on year t. The forecasting horizon is $h > 0$. Finally, the $\hat{\beta}_{i,h}$ is estimated by minimizing the following objective function:

$$\min_{\beta_{i,h}} \frac{1}{nT} \sum_{i}^{n} \sum_{t}^{T} (y_{i,t+h} - X'_{i,t}\phi_{i,h} - Z'_{i,t}\beta_{i,h}) + \frac{\lambda}{nT} \left[(1-\alpha) \left\| \beta_{i,h} \right\|_{\ell_1} + \alpha \left\| \beta_{i,h} \right\|_{\ell_2} \right] \tag{4}$$

The ℓ_1 and ℓ_2 are the elastic-net penalty, which is controlled by the two hyperparameters λ and α. I tune this parameters in a rolling window of train, validation and test set, which recursively the model is estimated on train data to minimize the mean squared forecast error on validation set. The test set is used to evaluate the model. I use two ways of estimation, one estimated by cross-section with all firms and another estimated by industry. For the cross-section estimation of the most predictive words I use $YEAR_{t-2}$ to train the model, while the industry estimation use the $YEAR_{t-5}$ to $YEAR_{t-2}$ as train data. The $YEAR_{t-1}$ is always used as validation set. This approach is similar to used by Frankel et al. (2016). I also perform a third model using a dimension reduction with the cross-section estimation.

Step 3 - Forecast Investment Growth

To forecast investment growth annually, I perform three different approach. In the first one I use the model estimated on previous step with a cross-section regression and forecast next year investment growth as equation (5) using $\hat{\alpha}$, $\hat{\phi}$ and $\hat{\beta}$ estimated in previous step with the $\mathbf{X}$ and $\mathbf{Z}$ predictors of year t. I also use each industry model and compute the next year expected investment growth also as equation (5), but here there is one model for each industry.

$$IG_{i,t+h|t} = \hat{\alpha} + \hat{\phi}_t X'_{i,t} + \hat{\beta}_{i,t} Z'_{i,t} \tag{5}$$

The last one approach is to get the high predictive words selected by equation (3) and define a $Z_t^* \subset Z_t$ for each year t, which Lima et al. (2020) calls time-varying dictionary. One difference from this work and the Lima et al. (2020) is that here I estimate each matrix Z_t^* by a cross-section regression. Even so, the dictionary tend to change over time. The Z_t^* is a high-dimensional matrix, and a dimensional reduction can improve forecast using this predictors. So bring insight from time-series approach of Lima et al. (2020), I pool the Z_{t-2}^*, Z_{t-1}^* and Z_t^* to define a a large matrix in which I estimate common factors by principal components. Than I take insight from Lima et al. (2020) on time-series domains and select the optimal number of factors via eigenvalue ratio approach of Ahn and Horenstein (2013). Then, I keep only the factors with p-value less than or equal to 0.01 in prediction equation applied on year $t-1$ as Bai and Ng (2008).

2.5 Monthly Text-based Measure

I also perform a monthly estimation of a text-based measure, which is more appropriate to do a cross-section statistical evaluation and to examine an economic value performance in time-series portfolio analysis. For the monthly estimation, I match monthly and year date as Hou et al. (2020) and George et al. (2018), that is that all accounting variables at month t is from the most recent fiscal year ending at least four months ago. One exception is the dRoe that is computed using earnings from the most recent announcement dates (item RDQ), and if not available, from the fiscal quarter ending at least four months ago. The word count from MD&A is from the most recent 10-k available for firm firm i at month t, using the SEC Edgars filing date.

With that, for each month I build three matrix to estimate the forecast model to predict investment growth from t to $t+h$ ($IG_{t+h|t}$). For example, to forecast investment growth from t to $t+12$ ($h=12$), I define a matrix $\mathbf{Y}$ with the most recent available investment growth

at month t (i.e. investment-to-assets change from $t-12$ to t), a matrix $\mathbf{X}$ with the most recent traditional predictors at month $t-12$ (e.g. Tobin's q, cash flow and change in return on equity from the fiscal year(quarter) ending at least six months(four months) ago) and, a large matrix $\mathbf{Z}$ for words in most recent MD&A available at month $t-12$.

The matrix $\mathbf{Z}$ with the word count is a high dimensional sparse matrix, and is not suitable to use as predictors in the traditional OLS regression. So I estimate a forecast model by using regularization method in a cross-section manner, which is similar to the approach of Han et al. (2020) that applied this idea to structured data. In analogous way I follow similar steps for text data.

Therefore, I implement a model to forecast the expected investment growth ($y_{i,t+h}$) in time h for each firm i, each month t I use a set of information available up to time t, so I estimate a cross-section model by using elastic net procedure as the linear prediction equation (6).

$$y_{i,t|t-s} = \phi_t X'_{i,t-s} + \beta_{i,t} Z'_{i,t-s} + \varepsilon_{i,t} \tag{6}$$

Where $h > 0$ is the forecasting horizon and $\hat{\beta}_{i,h}$ is estimated by minimizing the following objective function:

$$\min_{\beta_{i,h}} \frac{1}{nT} \sum_{i}^{n} \sum_{t}^{T} (y_{i,t+h} - X'_{i,t}\phi_{i,h} - Z'_{i,t}\beta_{i,h}) + \frac{\lambda}{nT} \left[(1-\alpha) \left\| \beta_{i,h} \right\|_{\ell_1} + \alpha \left\| \beta_{i,h} \right\|_{\ell_2} \right] \tag{7}$$

Where $X_{i,t}$ is the $p \times 1$ vector with p variables, and ℓ_1 and ℓ_2 are the elastic-net penalty, which is controlled by the two hyperparameters λ and α. The α bridges the gap between lasso ($\alpha = 1$) and ridge ($\alpha = 0$) regression and, the tuning parameter λ controls the overall strength of the penalty.

The Elastic Net estimation involves two non-negative hyperparameters, which imply in two well known regularizers as special cases. The LASSO case ($\alpha = 1$), which use absolute

value, or $\ell 1$, as parameter penalization. And the Ridge Regression case ($\alpha = 0$), which uses $\ell 2$ parameter penalizaion to draw all coefficients estimates closer to zero but does not impose exact zero anywhere. By generating linear models through both shrinkage and selection, Elastic Net seems to be suitable to my research problem, since I have a high dimensional sparse matrix as predictor.

For the monthly estimations, I set the tuning parameters with the intention of maximizing the prediction accuracy while maintaining the low computational intensity of the method. Than, I set $\alpha = 0.5$ and choose λ using the Hurvich and Tsai (1989) corrected version of the AIC (Akaike Information Criterion). Which is a similar approach used by Rapach and Zhou (2020), but here applied to text data.

By setting $\alpha = 0.5$, there is a stronger tendency for the model to select highly correlated predictors as a group (Hastie, Qian, & Tay, 2016; Rapach & Zhou, 2020). For the λ, I select via corrected AIC as Rapach and Zhou (2020) and Han et al. (2020). Despite the K-fold cross-validation be a popular way for tunning the parameters, setting the number of folds and their definitions can be extensively arbitrary, and the results can be sensitive to these decisions (Han et al., 2020). In addition, AIC procedure is more computationally scalable approach and, as doccumented by Flynn, Hurvich, and Simonoff (2013) and Taddy (2017), select λ via corrected AIC outperforms conventional five-fold cross-validations.

2.6 Performance Evaluation

For the yearly estimations, I compute the root mean squared forecast error (RMSFE) as equation (8), which compares from a conditional model (my propose) to that from the unconditional model (the benchmark). A similar approach is used by Lima et al. (2020) in the time-series domain, however here I apply the RMSE to pools prediction errors across firms and over time. As Gu, Kelly, and Xiu (2020), I perform assessment of each model by applying the out-of-sample evaluation measure into a panel-level.

$$RMSFE_j^h = \frac{\sqrt{\Sigma_{i=1}^{P}(IG_{t+h,i} - f_{t+h,i}^{j})^2}}{\sqrt{\Sigma_{i=1}^{P}(IG_{t+h,i} - f_{t+h,i})^2}} \tag{8}$$

where:

$IG_{t+h,i}$ = future realized investment growth from year t to $t+h$ for firm i;

$f^{j}_{t+h,i}$ = expected investment growth for $t+h$ of firm i predicted by elastic net model by using MD&A session of 10-K filings;

$f_{t+h,i}$ = expected investment growth for $t+h$ of firm i predicted by benchmark model. That is, classic Fama and MacBeth (1973) procedure as Hou et al. (2020).

For the monthly estimation, I follow Han et al. (2020) to evaluate models by computing the cross-sectional MSFE (mean square forecast errors) by specify in terms of deviations from the mean as equation (9).

$$MSFE_{\hat{f},t+h} = \frac{1}{N}\sum_{i=1}^{N}\left[(IG_{i,t+h} - \overline{IG}_{i,t+h}) - (\hat{f}_{i,t+h|t} - \overline{\hat{f}}_{i,t+h|t})\right]^2 fot \;\; t = 1, ..., T \tag{9}$$

where:

$IG_{t+h,i}$ = future realized investment growth from month t to $t+h$ for firm i;

$\hat{f}_{i,t+h|t}$ = expected investment growth for $t+h$ of firm i predicted by elastic net model by using MD&A session of 10-K filings;

$\overline{\hat{f}}_{i,t+h|t}$ = is the cross-sectional mean for $\hat{f}_{i,t+h|t}$;

$\overline{IG}_{i,t+h}$ = is the cross-sectional mean for $IG_{t+h,i}$.

That is as relevant metric for assessing cross-sectional forecasts because is concerned with relative expected growth across firms, in other words, I measure the cross-sectional differences in expected growth. For instance, consider the traditional MSFE as equation (10).

$$MSFE^{\dagger}_{\hat{f},t+h} = \frac{1}{N}\sum_{i=1}^{N}\left(IG_{i,t+h} - \hat{f}_{i,t+h|t}\right)^2 fot \;\; t = 1, ..., T \tag{10}$$

If the forecast if perfect (i.e. $IG_{i,t+h} = \hat{f}_{i,t+h|t}$ for $i = 1, ..., N$), then it is obvious that both MSFE measures in equations (9) and (10) are equal zero. However, if $\hat{f}_{i,t+h|t} = IG_{i,t+h} + c$ for

$i = 1, ..., N$, then by equation (10) the traditional $MSFE^{\dagger}_{\hat{f},t+h} = c^2$, oppositely, according to the cross-section MSFE used here, the $MSFE_{\hat{f},t+h} = 0$ by the equation (9).

Modified Diebold-Mariano for cross-section

As Gu et al. (2020), I addapt the Diebold and Mariano (1999) test from time-series domain, to perform the out-of-sample differences in cross-section predictive accuracy between two models. Specifically, to test forecast performance of model A versus B, I use the equation (11).

$$DM_{AB} = \frac{\bar{d}_{AB}}{\hat{\sigma}_{\bar{d}_{AB}}} \tag{11}$$

where:

$$d_{AB,t+h|t} = \frac{1}{N}\sum_{i=1}^{N}\left(\left(\hat{e}^{A}_{t+h|t}\right)^2 - \left(\hat{e}^{B}_{t+h|t}\right)^2\right) \tag{12}$$

The $\hat{e}^{A}_{t+h|t}$ and $\hat{e}^{B}_{t+h|t}$ are the cross-section prediction error at time t using each model. The $\bar{d}_{AB}$ and $\hat{\sigma}_{\bar{d}_{AB}}$ is the time-series mean and Newey and West (1994) standard error of $\hat{\sigma}_{\bar{d}_{AB,t}}$. So the modified Diebold-Mariano test is now based on a single time series $\bar{d}_{AB,t+h|t}$ and is more likely to satisfy the conditions needed for asymptotic normality, and than, gives appropriate p-values for test of model comparison (Gu et al., 2020).

Cross-section Forecast Encompassing

Han et al. (2020) propose a forecast encompassing test for comparing the information content of two competing cross-section forecasts. The test is based on Harvey, Leybourne, and Newbold (1998) from time-series domain. To compute the test, we perform the OLS regression as equation (13).

$$\hat{e}^{A}_{i,t+h|t} = \eta_t + \theta_t(\hat{e}^{A}_{i,t+h|t} - \hat{e}^{B}_{i,t+h|t}) + \epsilon_{i,t} \; for \; i = 1, ..., N; t = 1, ..., T, \tag{13}$$

where

$$\hat{e}^{k}_{i,t+h|t} = IG_{i,t+h} - \hat{f}_{i,t+h|t} \;\; for\; k = A, B. \tag{14}$$

Han et al. (2020) shows that estimate of θ_t, $\hat{\theta}_t$, in the equation (13) is identical to mimizes the month-t cross-sectional $MSFE^*$ of a forecast composite by two competing models as equation (15).

$$MSFE^*_t = \frac{1}{N}\sum_{i=1}^{N}\left[(IG_{i,t+h} - \overline{IG}_{i,t+h}) - (\hat{f}^*_{i,t+h|t} - \overline{\hat{f}}^*_{i,t+h|t})\right]^2 fot \;\; t = 1, ..., T \tag{15}$$

where

$$\hat{f}^*_{i,t+h|t} = (1-\zeta)\hat{f}^A_{i,t+h|t} + \zeta\hat{f}^B_{i,t+h|t} \;\; for\; i = 1, ..., N; t = 1, ..., T; 0 \leq \zeta \leq 1. \tag{16}$$

Finally using the Fama and MacBeth (1973) procedure, I take the time-series average of the monthly slope coefficient of equation (13) and test the null hypothesis that model A encompasses B ($\theta > 0$) and the null hypothesis that model B encompass A ($\theta < 1$). For this procedure, I compute the robust standard errors of Newey and West (1994) for $\{\hat{\theta}_t\}_{t=1}^{T}$ in the equation (17).

$$\hat{\theta} = \frac{1}{T}\sum_{t=1}^{T}\hat{\theta}_t \tag{17}$$

Modified Clark-West for cross-section nested model

One of the disadvantages of the last two tests is that they are not suitable for nested models, so I use the same procedure of Gu et al. (2020) and Han et al. (2020) to perform a modified Clark and West (2007) for cross-secion nested models. In other words, I compute the Clark and West (2007) on each cross-section as equation (18).

$$CW_{\hat{f},t+h} = (IG_{i,t+h} - \hat{f}^{benchmark}_{i,t+h|t})^2 - \left[(IG_{i,t+h} - \hat{f}^{text}_{i,t+h|t})^2 - (\hat{f}^{benchmark}_{i,t+h|t} - \hat{f}^{text}_{i,t+h|t})\right] \tag{18}$$

where:

$IG_{t+h,i}$ = future realized investment growth from month t to $t+h$ for firm i;

$\hat{f}^{text}_{i,t+h|t}$ = expected investment growth for $t+h$ of firm i predicted by elastic net model by using MD&A session of 10-K filings;

$\hat{f}^{text}_{i,t+h|t}$ = expected investment growth for $t+h$ of firm i predicted by benchmark model.

Then I take the time-series average as equation (19) to test the null hypothesis $CW_{\hat{f},t+h} \geq 0$ by using the robust standard errors of Newey and West (1994). In sum, the error differences are based on a single time series with little autocorrelation and is more possible to satisfy the mild regularity conditions needed for asymptotic normality, and in turn, gives appropriate p-values for comparison of the nested models (Gu et al., 2020). Although, any potential autocorrelation problem is mitigated by the Newey and West (1994) procedure.

$$CW = \frac{1}{T}\sum_{t=1}^{T} CW_t \tag{19}$$

2.7 Economic Value

Additionally, I analyze the performance of portfolios formed based on the proposed investment growth measure. So, in order to evaluate the economic implication of the cross-sectional out-of-sample investment growth forecasts, I construct long-short portfolios by sorting stocks according to their text-based investment growth measure. Precisely, at the end of each month, I sort stocks into equal-weighted quintiles based on their subsequent forecasted investment growth. I then construct a zero-investment portfolio that goes long (short) the highest (lowest) quintile.

3 Empirical Results

In this session, I show that the most predictive words are not always the most obvious and that they change according to the sector and over time, which shows how important it is to use a more flexible method to deal with a text-based forecast.

3.1 Models Estimated Recursively by Year

3.1.1 High predictive words

The Table 1 presents the most relevant words in the cross-section predictive model, the table displays the average coefficients ordered by the most positive (negative) value. The results are presented in the table by sub-sample (1994 to 2006 and 2007 to 2019) and full sample, which help to understand how the time-varying dictionary updates the most predictive words with the main objective of obtaining the best forecast. Another important insight from this table is that the coefficients are all very close to zero in the cross-section model. Although it still improves the forecast compared to the benchmark model, the textual model in the cross-section estimation seems to have difficulty to finding a strong pattern between MD&A and investment growth, probably due to the large variation in the firms financial reports with different characteristics, such as life cycle and industry.

[Table 1 about here.]

For better understand the estimation by industry, the Table 2 shows the high predictive words in four different industries: Health Care Equipment & Supplies (GICS 351010), Household Durables (GICS 252010), Containers & Packaging (GICS 151030) and Metals & Mining (GICS 151040). This results shows that in all industries the coefficients shows a larger value than in cross-section estimation, which inply that the approach used by Frankel et al. (2016) seems to estimate better models.

Some positive words seems to have a intuitive relation to investment growth such as the positive word "investing" in Containers & Packaging (GICS 151030) and the negative word "unrealized" for Household Durables (GICS 252010). However, as in Frankel et al. (2016) there is also counter intuitive words or with no clear relation such as the word "small" classified as positive in Health Care Equipment & Supplies (GICS 351010) and "procedures" classified as relevant word for both Household Durables (GICS 252010) and Metals & Mining (GICS 151040). See the high predictive words of others industries in Appendix **??**.

[Table 2 about here.]

Table 3 displays the high predictive words by life cycle. The words are ranked by the average coefficients of each one, and separated in positive and negative coefficients. The table shows the importance of a approach with no fixed dictionary, since most predictive words have no negative or positive connotation. However, there are some exception such as words that may charge a negative sentiment like “declines” in Introduction stage, “bad” in Shadec/Decline stage. And positive as well like “approvals” and "profitability" in Growth stage.

[Table 3 about here.]

3.1.2 Expected Investment Growth Forecasting Evaluation

The Table 4 present root mean squared forecast error computed as Equation (8). By these results, the combination of text regression with supervised machine learning to predict investment growth expectations from the MD&A section of 10-K filings leads to a better forecast. However, the first model shows a poor forecast for h = 1 and 2, which implies that using all firms to estimate the coefficient may not be a good approach when using words from financial reports as predictors.

[Table 4 about here.]

The second and third models do a better job, when grouping firms by sector or life cycle, there is a greater similarity between the reports or in the relationship between words and fundamentals. In addition to grouping by industry, the model has a better performance according to Frankel et al. (2016), grouping by life cycle leads to a model with a relative performance even higher than that of the industry.

3.1.3 Applying this text-based forecast to others fundamentals

For check flexibility of the text-based forecast procedure proposed in this work, I try to add text information for three different models of firm fundamentals, two of than is a different approach for investment growth, and the last one is a model to predict the popular return on equity (ROE), which is vastly useful for investment professionals.

First, I perform here the same analysis as the previous section, but instead the investment-to-assets change as used by Hou et al. (2020), now I use the CAPEX growth as Li and Wang (2018) and CAPEX-to-capital growth as used by George et al. (2018).

For the fist robustness check, I follow Li and Wang (2018) computing CAPEX growth in two steps. In the first step, I run the following annual cross-sectional predictive regression based on three predictors (Equation (20)). To reduce the impact of microcaps, the regression bellow is estimated by using weighted least squares with the market equity as the weights. Both the left- and right-hand side variables are winsorized each month at the 1% and 99% level.

$$E_{i,t}[IG_{LW}] = b_{0,t} + b_{MOM,t}MOM_{i,t-1} + b_{q,t}Q_{i,t-1} + b_{CF,t}CF_{i,t-1} + \epsilon_{i,t} \tag{20}$$

where:

$E_{i,t}[IG]$ = the growth rate of investment expenditure in the fiscal year ending in calendar year t ($IG = log(CAPEX_{i,t}/CAPEX_{i,t-1})$);

$MOM_{i,t-1}$ = the momentum cumulative stock returns over the past 12 months skipping one month before the end of last fiscal year;

$Q_{i,t-1}$ = the log of the market value of the firm divided by total assets in the fiscal year ending in calendar year $t-1$;

$CF_{i,t-1}$ = is the operating cash flow in the fiscal year ending in calendar year $t-1$ divided by lag total assets.

In the second step, compute the monthly EIG as the out-of-sample predicted value of investment growth from Equation (20) using the most up-to-date momentum, q and CF for each firm with the historical average of the cross-sectional regression coefficients $(b_{0,t}, b_{MOM,t}, b_{q,t}, b_{CF,t})$ estimated up to year t. Precisely, the accounting information as Q and CF are from fiscal year ending in calendar year t and the MOM (momentum) is the priour 2 to 12-month cumulative stock returns. I require a minimum of five years of regression coefficients to construct EIG in order to alleviate the impact of estimation errors. This proxy of investment plans used by Li

and Wang (2018) is used as the first benchmark in this robustness check, namely here, as the Expected Investment Growth of LW ($E[IG_{LW}]$).

I also test whether my text forecast procedure is able to improve the prediction of the model used by George et al. (2018), which measure investment growth as Liu, Whited, and Zhang (2009) by using the annual investment-to-capital (I/K), where investment (I) is capital expenditures (annual item CAPX) minus sales of property, plant and equipment (annual item SPPE, set to zero if missing); and capital (K) is net property, plant and equipment (annual item PPENT). Note that investment can be negative if firms downsize. Consequently, the simple ratio of the current year's I/K to the previous year's I/K can be negative even if investment is higher in the current year than in the previous year. To avoid this, we calculate the investment growth for fiscal year FY+1 (IG_{FY+1}) as Equation (21).

$$IG_{i,t} = \left[1 + \frac{I_t}{K_t}\right] / \left[1 + \left(\frac{I_{t-1}}{K_{t-1}}\right)\right] \tag{21}$$

The measure of George et al. (2018) has two main difference with the previous model (Equation (20)). First, the estimation of the parameters is monthly rather than annual. Second, to estimate the parameters used to forecast investment growth, George et al. (2018) used as dependent variable the CAPEX-to-capital change rather than just CAPEX growth. See Equation 22.

$$E_{i,t}[IG_{GHL}] = b_{0,t} + b_{ROE,t}ROE_{i,t-1} + b_{PTH,t}PTH_{i,t} + b_{PTL,t}PTL_{i,t} + \epsilon_{i,t} \tag{22}$$

where:

$E_{i,t+1}[IG_{GHL}]$ = the growth rate of investment-to-capital (as Equation 21) in the fiscal year ending in calendar year t;

$ROE_{i,t}$ = last available ROE, which is calculated by income before extraordinary items divided by two-year-lagged book equity;

$PTH_{i,t}$ = the ratio of current price to 12-month high price;

$PTL_{i,t}$ = the ratio of current price to 12-month low price.

Finally, to highlight the flexibility of the method proposed in this study, I apply the same approach of text regression and machine learning in order to use textual information from MD&A to predict others firms fundamentals, as return on equity (ROE). So I compete my text model with the Fama and MacBeth (1973) procedure used by George et al. (2018) as equation (23).

$$E_{i,t}[ROE] = b_{0,t} + b_{ROE,t}ROE_{i,t-1} + b_{PTH,t}PTH_{i,t} + b_{PTL,t}PTL_{i,t} + \epsilon_{i,t} \tag{23}$$

where:

$E_{i,t+1}[IG_{GHL}]$ = the growth rate of investment-to-capital (as Equation 21) in the fiscal year ending in calendar year t;

$ROE_{i,t-1}$ = last available ROE, which is calculated by income before extraordinary items divided by two-year-lagged book equity;

$PTH_{i,t}$ = the ratio of current price to 12-month high price;

$PTL_{i,t}$ = the ratio of current price to 12-month low price.

The table 5 shows the RMSFE of the models for alternative measures. For the prediction of alternative measures, the model was not able to be as efficient. Perhaps a way of combining Li and Wang (2018) and George et al. (2018) predictions with text prediction could yield better results. As for the prediction of other fundamentals, the text-based model proved to be much more efficient in predicting the ROE, indicating that the model can be flexible and applicable to other accounting fundamentals.

[Table 5 about here.]

3.2 Models Estimated Recursively by Month

[Table 6 about here.]

3.2.1 High predictive words

The Table 6 presents the top-25 high predictive words ranked by the average sign. The table also presents the words that is sentiment charged according Loughran and McDonald (2011) dictionary. In this high predictive words only three is sentiment charged, in other words, the approach of Lima et al. (2020) to not use a fixed dictionary seems to be suitable to cross-section forecast as well (in all words selected by the model, only 6.25% is sentiment charged). In addition, two important words to predict future investment growth is decrease and reduce, which can be associated increase in investment plans due to postpone projects, since this can be related to a reduction in a current asset.

3.2.2 Expected Investment Growth Forecasting Evaluation

To asses the accuracy of the monthly text-forecasts, Table 7 reports the time-series average of the monthly Diebold-Mariano statistic, R^2_{OOS} and Clark West test for nested models. Also show the time-series average of θ and the null hypothesis test for $\theta = 0$ and $\theta = 1$, from the encompass test. All statistics is computed using robust standard errors of Newey and West (1994).

[Table 7 about here.]

The Diebold-Mariano test shows that our model outperform the benchmark. By the R^2_{OOS} my model is 2.68% higher than classic model, which imply that words bring new set of information. The θ of 0.0784 shows that the benchmark model does not encompass the text model, which is confirmed by statistical test for null hypothesis that $\theta = 0$, and in contrast our model does not encompass the classic since the null hypothesis of $\theta = 1$ is rejected as well. Finally, the null hypothesis that the forecast error of my model is higher than the forecast error of benchmark is rejected by the test of Clark West, which account for difference in nested models.

3.3 Long-Short Portfolios Performance

To infer about the economic value of proposed forecast method, the economic value evaluation based on long-short portfolio performance are presented in Table 8. The table relates annualized mean, volatility and Sharpe ratio for each long-short portfolio. The portfolios go long (short) every month in stocks which the firm has the highest (lowest) investment growth forecast for the next fiscal year. The table shows the result for the value- and equal-weighting returns. The period is from 1996 to 2018, so is useful to compare the results with the wide market performance which has the lowest average return. However the volatility of the both portfolios are riskier than the wide market return, and only the equal weighting present a better Sharpe Ratio. Despite the poor performance of the value-weighting portfolio by the Sharpe ratio, the alpha of the Fama and French (2015) 5 factors model is positive and significant. For the equal-weighting portfolio, the performance in the period is better both by the Sharpe ratio (2.04) and the 5 factor annualized alpha (34.30).

[Table 8 about here.]

4 Conclusion

In this study, I propose a new measure of firm-level investment plans based on text data from MD&A (Management Discussion and Analysis) disclosure in 10-K filings. Specifically, I combine the idea of time varying dictionary of Lima et al. (2020) with the cross-section forecast procedure of Han et al. (2020), which is adapted here to text data.

The contribution of this work is twofold. First, I show that the words matters even to predict investment growth, which is empirically challenging to measure in the firm-level. In addition, by adapting the Han et al. (2020) procedure to text data, I contribute to the forecast literature that lacks to explore unstructured data in cross-section forecast. Second, I add to the investment literature by proposing to use machine learning tools and text data to predict investment plans, which I show that to measure including text-data generate more accurate predictions and better performance in long-short portfolios.

Following Frankel et al. (2016), I also try some variations of yearly estimations, including the estimation using all firms in each year, the estimation by industry, by life-cycle and the estimation using all firms with dimensional reduction using principal component analysis.

By this annual tests, I could conclude that estimate the coefficients by using all 10-K filings firms at once did not present a tolerable forecast, mainly in the short term. Possibly this result in the cross-section estimation occur due to the variability that exists between reports from different firms.

Therefore, separating firms into groups is a solution that leads to better forecasting. That is, words are important and machine learning models can lead to better prediction, but for that to separate firms by industry makes machine learning models find a stronger pattern between words and fundamentals. Another insight is that the results shows, according to common sense, that industry and life cycle are good ways to set the training sample. But in addition, I present new evidence that to predict expected growth in life-cycle investment appears to be more important than industry.

References

Ahn, S. C., & Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, *81*(3), 1203–1227.

Bai, J., & Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, *146*(2), 304–317.

Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of econometrics*, *138*(1), 291–311.

Cochrane, J. H. (1991). Production-based asset pricing and the link between stock returns and economic fluctuations. *The Journal of Finance*, *46*(1), 209–237.

Diebold, F., & Mariano, R. (1999). Comparing predictive accuracy', journal of business and economic statistics, 13 (3), july, 253-63. *INTERNATIONAL LIBRARY OF CRITICAL WRITINGS IN ECONOMICS*, *108*, 263–273.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, *116*(1), 1–22.

Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of political economy*, *81*(3), 607–636.

Flynn, C. J., Hurvich, C. M., & Simonoff, J. S. (2013). Efficiency for regularization parameter selection in penalized likelihood estimation of misspecified models. *Journal of the American Statistical Association*, *108*(503), 1031–1043.

Frankel, R., Jennings, J., & Lee, J. (2016). Using unstructured and qualitative disclosures to explain accruals. *Journal of Accounting and Economics*, *62*(2-3), 209–227.

Gennaioli, N., Ma, Y., & Shleifer, A. (2016). Expectations and investment. *NBER Macroeconomics Annual*, *30*(1), 379–431.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–74.

George, T. J., Hwang, C.-Y., & Li, Y. (2018). The 52-week high, q-theory, and the cross section of stock returns. *Journal of Financial Economics*, *128*(1), 148–163.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Han, Y., He, A., Rapach, D., & Zhou, G. (2020). Firm characteristics and expected stock returns. *Available at SSRN 3185335*.

Harvey, D. I., Leybourne, S. J., & Newbold, P. (1998). Tests for forecast encompassing. *Journal of Business & Economic Statistics*, *16*(2), 254–259.

Hastie, T., Qian, J., & Tay, K. (2016). *An introduction to glmnet.*

Hou, K., Mo, H., Xue, C., & Zhang, L. (2018). q5. *Working Paper*, 67.

Hou, K., Mo, H., Xue, C., & Zhang, L. (2020). An augmented q-factor model with expected growth. *Review of Finance*.

Hou, K., Xue, C., & Zhang, L. (2015). Digesting anomalies: An investment approach. *The Review of Financial Studies*, *28*(3), 650–705.

Hou, K., Xue, C., & Zhang, L. (2018). Replicating anomalies. *The Review of Financial Studies*.

Lamont, O. A. (2000). Investment plans and stock returns. *The Journal of Finance*, *55*(6), 2719–2745.

Li, J., & Wang, H. (2018). The Expected Investment Growth Premium. *Working Paper*.

Li, J., Wang, H., & Yu, J. (2020). Aggregate expected investment growth and stock market returns. *Journal of Monetary Economics*.

Lima, L. R., Godeiro, L. L., & Mohsin, M. (2020). Time-varying dictionary and the predictive power of fed minutes. *Computational Economics*, 1–33.

Lin, Q., & Lin, X. (2018). Expected investment and the cross-section of stock returns. *Economics Letters*, *172*, 43–49.

Liu, L. X., Whited, T. M., & Zhang, L. (2009). Investment-based expected stock returns. *Journal of Political Economy*, *117*(6), 1105–1139.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, *66*(1), 35–65.

Manela, A., & Moreira, A. (2017). News implied volatility and disaster concerns. *Journal of Financial Economics*, *123*(1), 137–162.

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, *31*(2), 87–106. Retrieved from `http://pubs.aeaweb.org/doi/10.1257/jep.31.2.87` doi: 10.1257/jep.31.2.87

Newey, W. K., & West, K. D. (1994). Automatic Lag Selection in Covariance Matrix Estimation. *The Review of Economic Studies*, *61*(4), 631–653. Retrieved from `https://academic.oup.com/restud/article-lookup/doi/10.2307/2297912` doi: 10.2307/2297912

Rapach, D. E., & Zhou, G. (2020). Time-series and cross-sectional stock return forecasting: New machine learning methods. *Machine Learning for Asset Management: New Developments and Financial Applications*, 1–33.

Taddy, M. (2017). One-step estimator paths for concave regularization. *Journal of Computational and Graphical Statistics*, *26*(3), 525–536.

Table 1: High predictive words in different periods

	1995-2006		2007-2019		All Sample	
	Positive words	coeff	Positive words	coeff	Positive words	coeff
1	compare december	0.02	dac	0.06	dac	0.06
2	interestbearing	0.02	consolidate statement	0.02	consolidate statement	0.02
3	entertainment	0.01	apollo	0.02	apollo	0.02
4	accounting principle	0.01	percent million	0.02	percent million	0.02
5	obtained	0.01	basel	0.01	compare december	0.02
6	acquire business	0.01	material cost	0.01	basel	0.01
7	electronics	0.01	portfolios	0.01	entertainment	0.01
8	follows	0.01	gas price	0.01	material cost	0.01
9	mariner health	0.01	loan facility	0.01	accounting principle	0.01
10	earning share	0.01	fiscal due	0.00	obtained	0.01
11	funded	0.01	annum	0.00	acquire business	0.01
12	work	0.01	increase percent	0.00	portfolios	0.01
13	recovery	0.01	risks	0.00	electronics	0.01
14	operation	0.01	commissions	0.00	follows	0.01
15	restructuring plan	0.01	oil gas	0.00	mariner health	0.01
	Negative words	coeff	Negative words	coeff	Negative words	coeff
1	earning	-0.03	clinical development	-0.03	earning	-0.03
2	revenue fiscal	-0.02	ownership product	-0.02	clinical development	-0.03
3	contract manufacturer	-0.02	mortgages	-0.01	revenue fiscal	-0.02
4	nonrecurring	-0.02	precious metal	-0.01	contract manufacturer	-0.02
5	opportunities	-0.01	cost sell	-0.01	ownership product	-0.02
6	severance	-0.01	month period	-0.01	nonrecurring	-0.02
7	solid waste	-0.01	period january	-0.01	opportunities	-0.01
8	gas property	-0.01	aggregate principal	-0.01	severance	-0.01
9	business combination	-0.01	supplementary data	-0.01	solid waste	-0.01
10	internet	-0.01	statement supplementary	-0.01	mortgages	-0.01
11	share common	-0.01	senior unsecured	-0.00	gas property	-0.01
12	telecommunications	-0.01	servicing	-0.00	business combination	-0.01
13	million share	-0.01	delaware basin	-0.00	internet	-0.01
14	six	-0.01	lenders	-0.00	share common	-0.01
15	wells	-0.01	december increase	-0.00	precious metal	-0.01

This table present average coefficients ordered by the most positive (negative) value of the words in the cross-section model, which is estimated using $YEAR_{t-2}$ as training data and $YEAR_{t-1}$ as validation data.

Table 2: High predictive words and phrases by Industry.

	GICS 351010		GICS 252010		GICS 151030		GICS 151040	
	Positive words	Coeff.	Positive words	Coeff.	Positive words	Coeff.	Positive words	Coeff.
1	interestearning	0.19	safety	0.06	providers	0.10	procedures	0.26
2	broadband	0.11	andor	0.04	delivered	0.08	discounts	0.13
3	tier	0.09	solutions	0.03	audit	0.08	county	0.12
4	tenant	0.07	external	0.02	small	0.08	networks	0.10
5	charter	0.05	candidates	0.02	investing	0.07	barrel	0.10
6	refined	0.04	amortized	0.02	entertainment	0.07	interestbearing	0.08
7	small	0.02	agency	0.01	increasing	0.05	reference	0.07
8	monthly	0.02	phases	0.01	auto	0.04	billing	0.07
9	commodity	0.01	employment	0.01	director	0.04	conditional	0.05
10	vessels	0.01	material adverse	0.01	radio	0.04	leasing	0.04
	Negative words	Coeff.	Negative words	Coeff.	Negative words	Coeff.	Negative words	Coeff.
1	derived	-0.08	depletion	-0.18	observable	-0.23	satellite	-0.14
2	practices	-0.03	managements	-0.06	week	-0.10	patents	-0.14
3	membership	-0.03	procedures	-0.05	whether	-0.06	annuity	-0.07
4	adverse effect	-0.03	carried	-0.04	expectations	-0.06	guaranty	-0.07
5	advisory	-0.02	auto	-0.03	branch	-0.05	online	-0.06
6	franchise	-0.02	forth	-0.02	salaries	-0.04	generating	-0.04
7	refinery	-0.02	hotels	-0.02	collateral	-0.04	partnerships	-0.04
8	warrant	-0.02	materially	-0.01	depend	-0.04	gathering	-0.04
9	stores	-0.02	unrealized	-0.01	supplies	-0.03	predecessor	-0.04
10	branch	-0.02	inprocess	-0.01	unpaid	-0.03	weeks	-0.04

This table present average coefficients ordered by the most positive (negative) value of the words in the model estimated by industry, which use the $YEAR_{t-5}$ to $YEAR_{t-2}$ for each industry as train data and, the $YEAR_{t-1}$ is as the validation set. See the Appendix **??** for all industries.

Table 3: High predictive words and phrases by Life Cycle.

	Introduction		Growth		Mature		Shadec/Decline	
	Positive Words	coeff	Positive Words	coeff	Positive Words	coeff	Positive Words	coeff
1	interestearning	0.090	nonperforming	0.297	institution	0.049	casino	0.227
2	farmer	0.084	gap	0.270	noninterest	0.019	gaming	0.227
3	sensitivity	0.047	redevelopment	0.114	cement	0.018	premiums	0.119
4	duke	0.045	lae	0.103	percent million	0.017	served	0.088
5	compare december	0.027	riskbased	0.084	mortgages	0.014	mortgages	0.061
6	ongoing	0.026	aig	0.064	initial	0.012	interestearning	0.058
7	reflects	0.025	anticipate	0.047	indenture	0.009	electricity	0.042
8	electricity	0.025	mortgages	0.040	quoted	0.008	foreclosure	0.038
9	collateralized	0.024	approvals	0.039	trust	0.006	consists	0.037
10	mexico	0.024	profitability	0.036	unsecured	0.006	grade	0.037
	Negative Words	Coeff.	Negative Words	Coeff.	Negative Words	Coeff.	Negative Words	Coeff.
1	accruing	-0.113	interestbearing	-0.097	video game	-0.014	registrants	-0.143
2	central	-0.049	bankruptcy	-0.089	absolute	-0.013	nasdaq	-0.082
3	effectively	-0.044	noncovered	-0.029	foreclosure	-0.010	reit	-0.050
4	otherthantemporary	-0.036	reit	-0.026	certificates	-0.006	bad	-0.044
5	substantially	-0.035	accident	-0.023	impaired	-0.005	generating	-0.044
6	commission	-0.034	business acquisition	-0.017	order	-0.005	annuity	-0.040
7	declines	-0.034	tobacco	-0.015	rate note	-0.005	requires	-0.031
8	agency	-0.029	annuity	-0.014	clo	-0.004	policy	-0.023
9	currencies	-0.029	support service	-0.012	organic revenue	-0.004	control	-0.021
10	portfolio	-0.022	unrecognized	-0.011	auction rate	-0.004	lien	-0.018

This table present average coefficients ordered by the most positive (negative) value of the words in the model estimated by life-cycle, which use the $YEAR_{t-5}$ to $YEAR_{t-2}$ for each life-cycle classification as train data and, the $YEAR_{t-1}$ is as the validation set.

Table 4: RMSFE relative to the benchmark

	$h=1$	$h=2$	$h=3$
$EIG_{fixed-dictionary}$	0.965**	0.974*	0.882***
$EIG_{cross-section}$	0.955**	0.957*	0.873***
$EIG_{by-industry}$	0.920***	0.928***	0.897***
$EIG_{life-cycle}$	0.802***	0.829***	0.765***
$EIG_{dimension-reduction}$	0.922***	0.906***	0.892***

This table present the root mean squared forecast error (RMSFE), computed as ($RMSFE_j^h = \sqrt{\Sigma_{i=1}^{P}(IG_{t+h,i} - \hat{f}_{t+h,i}^{model_j})^2} / \sqrt{\Sigma_{i=1}^{P}(IG_{t+h,i} - \hat{f}_{t+h,i}^{benchmark_j})^2}$) of pools prediction errors across firms and over time in a panel-level. The subscripts *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

Table 5: RMSFE of alternative fundamentals relative to respective benchmark.

	cross-section	by industry	PCA
$E[IG_{i,t+1}]^{LWI}$	1.071	1.070	1.059
$E[IG_{i,t+1}]^{GHL}$	-	1.033	1.026
$E[ROE_{i,t+1}]$	0.965	0.971	0.975

This table present the root mean squared forecast error (RMSFE), computed as ($RMSFE_j^h = \sqrt{\Sigma_{i=1}^{P}(IG_{t+h,i} - \hat{f}_{t+h,i}^{model_j})^2} / \sqrt{\Sigma_{i=1}^{P}(IG_{t+h,i} - \hat{f}_{t+h,i}^{benchmark_j})^2}$) of pools prediction errors across firms and over time in a panel-level. The subscripts *, **, and *** indicate significance at the 10%, 5%, and 1% levels.

Table 6: Top 25 high predictive words on monthly estimation.

	variable	coef	sentiment	variable	coef	sentiment
1	decrease	3.68754		access service	-0.12242	
2	reduce	0.21439		electronic security	-0.12241	
3	license fee	0.09374		june	-0.10596	
4	reduction	0.07703		cable operator	-0.10388	
5	supply chain	0.05289		solid waste	-0.10165	
6	source	0.05045		fiber optic	-0.07430	
7	total net	0.04705		acquisition	-0.05921	
8	companys	0.03437		loan agreement	-0.05814	
9	insurance	0.03276		goodwill	-0.05634	
10	thousand	0.03088		vision system	-0.05619	
11	patent	0.03060		machine vision	-0.05353	
12	series prefer	0.02404		amortization	-0.05325	
13	company genta	0.02265		inprocess	-0.05228	
14	genta jago	0.01683		semiconductor	-0.05224	
15	care	0.01365		offer	-0.05061	
16	medical	0.01312		share series	-0.04585	
17	research	0.01156		avisof energy	-0.04411	
18	rb falcon	0.01050		public	-0.04054	
19	fda	0.00982		assurance	-0.04020	
20	secure note	0.00837		avisof utility	-0.03907	
21	collaborative	0.00501	positive	convertible	-0.03816	
22	termination	0.00498	negative	cost service	-0.03756	
23	discontinue	0.00306	negative	technology	-0.03566	
24	institution	0.00306		senior note	-0.03564	
25	yearend	0.00009		system	-0.03468	
26	web	-0.00232		product revenue	-0.03395	
27	wireless	-0.00558		placement	-0.03130	
28	assume	-0.00615	uncertainty	warrant	-0.02921	
29	absolute	-0.00708		july	-0.02808	
30	drill	-0.00839		financial institution	-0.02747	

This table present average coefficients ordered by the most positive (negative) value of the words in the cross-section model estimated monthly. The third column exhibit if a specific word is sentiment charged using Loughran and McDonald (2011) dictionary.

Table 7: Forecast Evaluation for Monthly Estimated Models

Diebold-Mariano	0.0243*
R^2_{OOS}	2.68%**
Encompass Test (θ)	0.0784
t-statistic ($\theta = 0$)	3.05***
t-statistic ($\theta = 1$)	35.87***
Clark West	0.0426***

This table present time-series average of cross-sectional evaluation measures computed each month. R^2 out-of-sample (R^2_{OOS}) is computed as $1 - (MSFE^{text}/MSFE^{benchmark})$. The table also present θ estimation from the encompass test, the time-series average of θ and the null hypothesis test for $\theta = 0$ and $\theta = 1$. All statistics is computed using robust standard errors of Newey and West (1994).

Table 8: Economic Value - Period 1996 to 2018

Panel A	Market	Value weighting	Equal weighting
Annualized Mean (%)	12.67	13.95	52.04
Ann. Volatility (%)	15.21	21.56	25.39
Ann. Sharpe Ratio	0.83	0.65	2.04

Panel B		Value weighting	Equal weighting
Annualized α (%)		10.65*	34.30***
MKT		0.132	0.188
SMB		-0.342	0.005
HML		-0.482	-0.305
RMW		-0.254	-0.373
CMA		1.375***	1.140***

The table reports annualized summary statistics for long-short portfolios constructed from out-of-sample forecasts of cross-sectional investment growth based on the MD&A. At the end of each month, I sort all available stocks into quintiles according to their forecasted investment growth for the next fiscal year. The long-short portfolio goes long (short) the fifth (first) quintile. The quintiles for the long-short portfolios are value (equal) weighted according to market capitalization. Market return in Panel A is the CRSP value-weighted market portfolio return minus the risk-free return. The subscripts *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively, according to t-statistics based Newey and West (1994) standard errors.