Measuring non-GAAP Earnings Quality Using Qualitative Disclosure: A Machine-learning Approach

Jack Wang

April 27, 2022

Abstract

I use machine learning methods to examine whether the qualitative information contained in 8-K earnings press releases is useful to assess firms' non-GAAP earnings quality. I train my model with 8-K earnings press release textual information and firm specific earnings response coefficient (ERC) as the measure of non-GAAP earnings quality ("the quantitative measure"). The trained model can then be used to generate a non-GAAP earnings quality score for each disclosing firm based on the qualitative information ("the qualitative measure"). I expect to document that the higher value of the qualitative measure is positively associated with more transitory non-GAAP exclusions.

1 Introduction

One of the most important implication of accounting earnings is for evaluating firm's performance. However, the accounting earnings calculated based on the General Accepted Accounting Principles (GAAP) is argued to be less informative for evaluation purpose because it contains transitory items that are not relevant with firm's core performance as oppose to the non-GAAP measure of earnings, from which managers exclude all the impacts of transitory items on accounting earnings. While tons of research has been done to examine the determinants and consequences of non-GAAP earnings and exclusions, the extant research mainly focuses on quantitative aspects of non-GAAP earnings. Qualitative disclosure, which comprises the majority of financial disclosure, can also provide useful and relevant information to stakeholders. Prior research has documented that firm qualitative disclosures can provide incremental information about firm performance and are associated with future stock market return (Insert citations). Therefore, in this paper, I am going to examine whether the earnings press release containing non-GAAP earnings provide incremental qualitative information that can be used to assess the quality of non-GAAP earnings.

In a recent research, Chen et al. (2021) introduced a measure of non-GAAP earnings quality measure based on qualitative characteristics of firm earnings press release. They hand-collect earnings press release with non-GAAP disclosures for SP 500 firms and assign score to each observation based on researchers judgements on whether each disclosure meet certain disclosure requirements. However,relying on researchers judgements may introduce unwanted bias and measurement errors for their disclosure quality scores and hand-collecting data limits their sample size. The advanced computing technology enables me to tackle these issues with more efficient and reliable methods. I scrape all the earnings press releases with non-GAAP disclosures from SEC EDGAR using Python. Then I use "tm" package in R to do the preprocess on the textual information, which returns a corpus in document-term matrix (DTM) format. This DTM, which is the independent variables in the later step machine learning model, contains all the qualitative information in earnings press releases. The purpose of this paper is to use the qualitative information to evaluate firm's non-GAAP disclosure quality. Currently in accounting research, ERC can be used to assess the quality of disclosure. Therefore, I use the firm specific ERC in the machine learning model as dependent variable to train and validate the model. The trained model is expected to generate a disclosure quality score of non-GAAP using the qualitative information in earnings press release. This research may contribute to the literature from several aspects. First, it provide additional evidence that qualitative information of non-GAAP disclosures contains information that can be used to evaluate non-GAAP earnings quality. While prior

research mainly focuses on the quantitative attributes of non-GAAP disclosures, the information usefulness qualitative non-GAAP disclosure is left under studied. Second, compared to the research that uses manually processed data, my research, using machine learning methods to process all the textual information, is less likely to be affected by researchers' subjectivity. Lastly, using machine learning methods also enables the analysis to incorporate more comprehensive information in the qualitative disclosure.

2 Literature Review

3 Data

The final training and testing data will contain firm specific ERC as the dependent variable and the counts of words in earnings press releases as the independent variables.

3.1 Data for calculating ERC

ERC is the coefficient on the unexpected non-GAAP earnings when we regress accumulated abnormal return (CAR) of the accounting period on the unexpected non-GAAP earnings. To calculate CAR, I use the financial data from Compustat and the stock market return data from CRSP dataset. For unexpected earnings calculation, the real non-GAAP earnings data is from Bentley et al. (2018) and the expected non-GAAP earnings are from I/B/E/S analysts forecast consensus. After deleting observations that do not have enough data to calculate ERC, I got the firm specific ERC for 10220 firm-year observation.

3.2 Earnings Press Releases Data

The textual information that is used as the input to the model is from the earnings press releases that include non-GAAP disclosure. I use python code to scrape the original earnings press releases from SEC EDGAR website. The whole textual information of each earnings press release is then put into one cell of the earnings press releases column of the corresponding firm-year observation in the data frame. In total, I got over 45,000 firm-year observations. However, since the original dataset, which contains the whole textual disclosures, is almost 1 GB, I can not process it on my computer due the the RAM limit of my personal computer. I randomly select 2,500 observations from the original data for further analysis.

Before the data is ready for machine learning training, I preprocess the selected sample using "tm" package in R. The pr column is the vector contains all the earnings press release textual information is transferred into a corpus data frame first. Then all the words are converted to lower case and are stemmed to the common root. All the white spaces, punctuation, stop words and numbers are removed. The processed data will then be used for further analysis.

4 Methods

4.1 To Calculate ERC

$$CAR = \alpha_i + \beta_i UE + \epsilon_i \quad (1)$$

Following prior research (Insert citation here!) I use the above model to calculate the firm specific ERC ($_i$). The model is estimated by firm-year using time series quarterly data of 20 quarters before the fiscal year end of t. CAR is the accumulated abnormal return, which is the sum of the abnormal return of each month in quarter q. The coefficients that are needed to calculate the abnormal return are derived from market model using monthly return of the past 24 months before the beginning of the quarter q. UE is the unexpected non-GAAP earnings of quarter q calculated on a per share basis, which is the difference between the real non-GAAP earnings and expected non-GAAP earnings. I use the quarterly managerial non-GAAP earnings per share (EPS) from Bentley et al. (2018) as the real non-GAAP earnings and the most recent analysts' consensus non-GAAP EPS estimate as the expected non-GAAP earnings.

4.2 Using sLDA to Estimate the Score

I use the supervised latent Dirichlet allocation (sLDA) model to estimate the non-GAAP disclosure quality based on qualitative input. sLDA chooses latent topics that are associated with a dependent variable by grouping phrases based on the probability of the phrases co-occurring within disclosures (Blei and McAuliffe 2007). Therefore, sLDA is more likely to identify the importance of groups of words and phrases when explaining a dependent variable. The lda package in R can be used to conduct the sLDA analysis. Specifically, the function slda.em is used for the analysis, which requires multiple inputs. The "documents" requires a list input, which in my case will be the preprocessed data. "K" is the number of topics, which can be obtained through running unsupervised LDA model. "vocab" is character vector including the vocabulary words used in documents, which can be got from the preprocessed corpus."params" is the initial values for the regression parameters, which I need to randomly assign. "annotations" is the outcome variable, which in my case would be the firm specific ERC. The results of sLDA analysis contain coefficients on each of topic identified by the model. These coefficients indicate to what extent each topic is associated with the outcome variable. However, obtaining topics is not the focus of this research. The predictions of outcome variable based on these topics is the interest of this research. The slda.predict function will give the estimated outcome, which is the non-GAAP earnings quality score based on qualitative information in earnings press release.

5 Findings

6 Conclusion

References