MACHINE LEARNING USING INSTRUMENTS FOR TEXT SELECTION: PREDICTING INNOVATION PERFORMANCE

[1]KIAN-GUAN LIM, [2]MICHELLE S.J. LIM

Singapore Management University
E-mail: kgl@smu.edu.sg

Abstract – In machine learning we utilize the idea of employing instrumental variable such as patent records to train the texts. Patent records are highly correlated with R&D expenditures, but are not necessarily correlated with performance residuals not linked to R&D. Thus, using instrumental patent records to train word counts of selected texts to serve as a proxy for firm R&D expenditure, we show that the texts and associated word counts provide effective prediction of firm innovation performances such as firm market value and total sales growth.

Keywords - Machine Learning; R&D Reporting; Textual Analyses; Firm Innovation

I. INTRODUCTION

This is a study on how machine learning methods in textual analyses can be improved by carefully selective training data sets that include instrumental variable set that avoids data snooping bias. It is also a study on how machine learning in textual analyses can profoundly help discover innovation activities and lead to a better assessment of a firm's innovation performance when R&D accounting information is missing. For more than half a century, corporate research and development (R&D) has been considered a primary measure of innovation in a firm (see Lerner and Wulf, 2007). R&D expenditures provide tangible long-term benefits to a firm (see Kothari et al., 2002). However, the accounting classification of R&D outlays is often within the discretion of a manager, resulting in ambiguous R&D expense reporting. Koh and Reeb (2015) also showed that there were firms, with missing R&D or zero R&D expenditures, but had filed and received patents on innovations.SomestudiesreplacedthemissingR&Dvalueswitheitherindustryaverage R&D or historical R&D values. In empirical research, the use of such ad hoc methods to manage missing R&D expenditures may provide misleading interpretations of firms' R&D performances.

To overcome the missing data problem in R&D expenditures, we harness the power of natural language processing and online availability of Securities and Exchange Commission (SEC) filings. We harvest textual revelations including new product announcements, R&D and other innovations, and qualitative but not quantitative suggestions of innovative progresses in the firms' SEC filings. The SEC filings are taken seriously as firms that are caught misrepresenting or hiding major developments may be subject to penalties by the SEC. Hence, textual analyses carry promises of revelation of innovations whereas annual accounts may not show R&D due to the discretionary accounting. Loughran and Mc- Donald (2016) is a recent survey of textual analysis in accounting and finance–textual analysis indeed has been applied to voluntary disclosures such as Allee and DeAngelis (2015) and others.

In machine learning, predicting some data sequence is often done using partitioning of the data into a prior training set and a subsequent test set. However, this modus operandi sometimes has problems when the predictors are texts that are selected from a wide lexicon without prior restrictions. This is because training can produce overfitting predictors or too large a set of texts such that the testing or actual prediction result becomes poor. This is similar to the endogeneity explanatory variable situation in linear regression with pre-defined variables. Firm R&D expenditures are known to be good predictors of firm innovation performance, but R&D expenses are sometimes not reported or reported in a discretionary manner that could bias any prediction of firm innovation performances. We use text variables to replace R&D expenditures in prediction when the latter are not available. Since we should not train the text variables using strategic and sometimes missing R&D expenses, and also not use actual dependent variable of performance measures to prevent data snooping on text variables, we utilize the idea of employing instrumental variable such as patent records to train the texts. Patent records are highly correlated with R&D expenditures, but are not necessarily correlated with performance residuals not linked to R&D.

We employ a smaller set of published patent data from 1994 to 2009 (see Kogan etal., 2017) (later data are not available) to train the selection of the root words and their chains to be able to significantly predict the subsequent year's patent numbers. Just as R&D expenditures may be considered a primary measure of firm innovation, patent data have also been widely regarded as indicators of technological change and innovation performance. Riitta (2000) and Zvi (1990) showed how patents were strongly related to R&D across firms and with innovation performances such as market value and sales. To perform the textual analysis, we firstly identify words that are commonly used to describe patent (unavoidably some are connected in the same sources with “innovation”) in accounting and financial contexts. Next, we parse the SEC filings for the occurrences of these words and record the frequency of the various word counts. This selection of textual word chains via the instrument of patent records ensures that there are no data snooping bias when weusethelistoftrainedwordstopredictfirminnovationperformancesintheabsenceof R&D expense records. Finally, we use the selected word chains in panel regressions to predict innovation performances realized in firms’ market values and in firms’ total sales activities, using data from 1994 to2017.

II. TEXTUAL ANALYSIS AND INNOVATION PREDICTION

The data used in this study comprise textual data collected from the U.S. SEC fil- ings, U.S. patent data downloaded from Noah Stoffman’s website1, and financial data sourced from the Compustat database. We download all the SEC filings from the Soft- ware Repository for Accounting and Finance (SRAF) website (https://sraf.nd.edu/) from 1994 to 2017. The1, 029, 963 SEC filings obtained from SRAF include the annual filings, 10-Ks, and quarterly filings, 10-Qs, as well as their variants. These text files exclude extraneous materials, such as tables and exhibits, that are more likely to contain template language and are deemed less meaningful in textual analysis (see Loughran and Mcdon- ald,2011).

This data was collected for the paper: Kogan, L., Papanikolaou, D., Seru, A. and Stoff- man, N., 2017. Technological innovation, resource allocation, and growth. Quarterly Journal of Economics, 132(2), pp.665-712. See SSRN version: http://ssrn.com/abstract=2193068 and also https://iu.app.box.com/v/patents. A detailed description of the data is provided in the paper.

\

Predictive panel regressions in Eq.(1) of log number of filed patents of a firm are run on respective log of word counts and annual year-end firm data of size, leverage, liquidity, and age as control variables. Patent variable is natural log of the annual number of patents filed from 1994 to 2009. Size is natural log of total assets. Leverage is natural log of (Total debt/Total equity). Liquidity is natural log of (Current assets/Current liabilities), and Age is the natural log of firm age in years. The words trained in the regression include *research*, *innovate*, *invent*, *pioneer*, and *spear head*. The data sample contains firms with at least one patent record but the firms need not have non-zero $R&D. Words are specifically associated with the mentioned firm. For example, the variable *Word* (*Research*) is the natural log of the number of words in *research* and its ot her root words counted in the U.S. SEC filings on all the firms in the sample. Panel data sample are annual from 1994 to 2009. Random effects are rejected using Hausman’s test at 5% significance level.

$$Patent_{i,t} = \beta_0 + \beta_1\ Word(Research) + \beta_2 Word(Innovate) + \beta_3 Word(Invent) + \beta_4 Word(Pioneer) + \beta_5 Word(Spearhead) + \beta_6 Size_{t-1} + \beta_7 Leverage_{t-1} + \beta_8 Liquidity_{t-1} + \beta_9 Age_{t-1} + s_{i,t} \quad (1)$$

For the controls, leverage has negative impact on patent. This shows that when firms are more highly in debt, they are more wary of incurring higher expenses and thus have less patents as outcomes. Liquidity does not have predictive effects though there appears to have contemporaneous negative associations.

From the panel regression we selected trained words of *Research*, *Innovate*, and *Invent*, and their roots, as they are significant in the regression in Eq.(1). We form *Word* as the log of the total counts for all the three words and their roots e.g. *innovation*, *inventing*, etc. We continue the investigation by using the entire sample of 1994 through 2017 data to examine the ability of the trained *Word* to predict firm innovation performances. The context of this prediction is in the case when $R&D expenses are not reported or which appear as zero entries in the accounts. We employ the following panel regression in Eq.(2).

Assessing the contribution of R&D investment to firm performance has been the interest of many studies. Specifically, the innovative capability of a firm may enhance its ability to offer valuable, rare, inimitable and differentiated products as well as services, result- ing in better financial performance (see Zahra et al., 2000). However, the production of innovative outputs requires significant investment of resources (see Hitt, Hoskisson, and Kim, 1997). Thus the firm’s performance as a result of innovations through R&D could be measured by market value normalized by cost proxied by sales. The selected texts can be used to predict the two aspects of firm innovations performance based on growth and on market value per unit sale.

$$Performance_{i,t} = \beta_0 + \beta_1 Word + \beta_2 Size_{t-1} + \beta_3 Leverage_{t-1} + \beta_4 Liquidity_{t-1} + \beta_5 Age_{t-1} + s_{i,t} \quad (2)$$

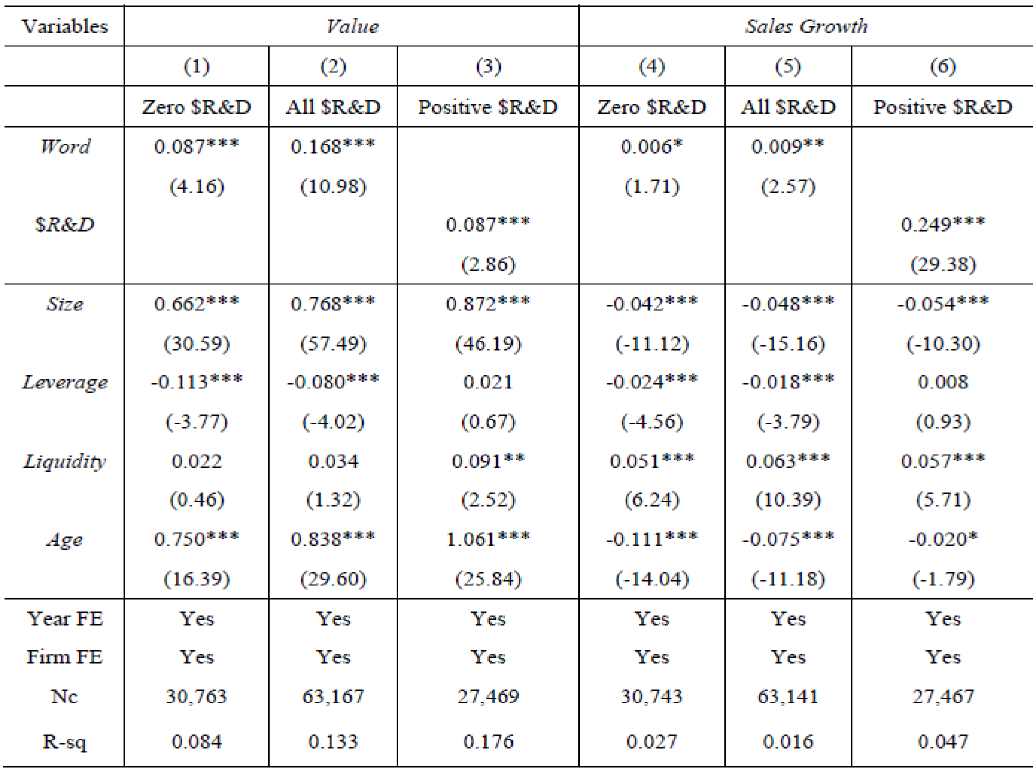
where *Word* is log of count of all words containing *Research*, *Innovate*, and *Invent*. Panel regression model is employed on two dependent variables: *Value* and *Sales Growth*. These variables are suitable measures of innovation performance. Market Value is the share price multiplied by the number of shares outstanding, Value is natural log of (Marketvalue/Totalsales).SalesGrowthisnaturallogoftheannualpercentagegrowthinfirm sales, The regression results are reported in Table 1below.

In Table 1, columns 1 and 4 show panel regression results using the sample of firms with zero $R&D. This addresses directly our intention of showing if textual analyses can predict innovation outcomes in the absence of R&D data. However, to provide a robust comparison we add panel regressions in columns 2 and 5 using the sample that includes also firms with positive as well as zero$ R&D.As innovation activity takes time to show performance, we apply a one-year lag (see Kafouros et al., 2018). The lag on the right-hand side also gives the panel regression a truly predictive framework, allowing the co-integrated stationary process to be unbiased in estimation.

As for the control variables, their coefficient estimates are consistent with the findings in the earlier sub-section on patent regression. Larger firms are associated with more patents and thus innovations (see Cohen et al., 2000). This results in positive impact on the innovation performance of *Value*.Highly leveraged firms may not have the resources or discretion to pursue innovative activities (see Zahra et al., 2000). Thus the firm's *Value* and *Sales Growth* are both negatively affected by increasing leverage. On the other hand, firms with higher liquidity may have more resources to support innovative activities (seeTanandPeng,2003). This results in positive impact on *Value* and *Sales Growth*. Younger firms may have more patents as inertia and sunk costs may deter older firms from investing in innovative activities (see Zahra et al., 2000). Thus *Age* has a negative coefficient in *Sales Growth*. However, *Age* and survival bias would imply longer and successful survival imply higher market value as seen in Table 1. We perform the Hausman test. The test rejects the random model in favour of the within-model or fixed effects model.

III. CONCLUSION

The empirical results in this paper suggest the feasibility and usefulness of employing textual analyses to discover innovation activities and provide predictive assessments of a firm's innovation performance such as market value and sales growth. The use of word counts as a proxy for firm innovation increases the number of firms that potentially harbor innovativeactivities.Thisisanadvantageoverusingjustt heR&Ddatawhichmaycurtail the number of firms and put firms using zero discretionary R&D reporting outside the probe for innovations. The strong positive significance of the prediction of *Word* for normalized market value as in *Value* and also positive significance of the prediction of *Sales Growth* show without doubt that textual analyses is key to identifying innovation activities and predicting value and growth in firms. Although the predicting variable *Word*isinitselfnotacardinalvariable,itscountiscardinal, andthepositiverelationship of the count (or log of count) to the quantity of value and growth indicates that more reporting and more frequent statements of such key root words as *Research*, *Innovate*, and *Invent* actually predict firm innovations performance. Moreover, to ensure that the *Word* replaces $R&D when it is absent, $R&D when present should also be able to predict the innovation performances of *Value* and *Sales Growth*.



Variables	*Value*			*Sales Growth*		
	(1)	(2)	(3)	(4)	(5)	(6)
	Zero $R&D	All $R&D	Positive $R&D	Zero $R&D	All $R&D	Positive $R&D
Word	0.087***	0.168***		0.006*	0.009**	
	(4.16)	(10.98)		(1.71)	(2.57)	
$R&D			0.087***			0.249***
			(2.86)			(29.38)
Size	0.662***	0.768***	0.872***	-0.042***	-0.048***	-0.054***
	(30.59)	(57.49)	(46.19)	(-11.12)	(-15.16)	(-10.30)
Leverage	-0.113***	-0.080***	0.021	-0.024***	-0.018***	0.008
	(-3.77)	(-4.02)	(0.67)	(-4.56)	(-3.79)	(0.93)
Liquidity	0.022	0.034	0.091**	0.051***	0.063***	0.057***
	(0.46)	(1.32)	(2.52)	(6.24)	(10.39)	(5.71)
Age	0.750***	0.838***	1.061***	-0.111***	-0.075***	-0.020*
	(16.39)	(29.60)	(25.84)	(-14.04)	(-11.18)	(-1.79)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Nc	30,763	63,167	27,469	30,743	63,141	27,467
R-sq	0.084	0.133	0.176	0.027	0.016	0.047

Table 1: Predicting firm innovation performances with textual analyses.

Predictive panel regressions of a firm's Value at *t* and also Sales Growth at *t* on respective word counts and annual year-end firm data of size, leverage, liquidity, and age as control variables at $t-1$. Words are specifically associated with the mentioned firm. *Word* is log of count of all words containing *Research*, *Innovate*, and *Invent*. Panel data sample are annual from 1994 to 2017. Panel regressions are with fixed effects (FE) on year and firm. Random effects are rejected using Hausman's test at 5% significance level. Columns 1 and 4 show panel regression results using the sample of firms with zero $ R&D. Columns 2 and 5 show panel regression results using the sample that includes also firms with positive as well as zero$ R&D. Columns 3 and 6 show regressions using $ R&D as explanatory variable in place of *Word*.

Notes: ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. The t-values are in parentheses. Number of firm-years used in the unbalanced panel regression. Number of sample points in columns 2 and 5 are greater than in columns 3 and 6 because the former include firm-years where $R&D can be both positive or zero.

REFERENCE

[1] Allee, K., and M. DeAngelis, (2015), "The Structure of Voluntary Disclosure Narra- tives: Evidence from Tone Dispersion", Journal of Accounting Research, 53, 241- 274.

[2] Cohen, W.M., R.R. Nelson, and J.P. Walsh, (2000), "Protecting their intellectual as- sets: Appropriability conditions and why U.S. manufacturing firms patent (ornot)", National Bureau of Economic Research, Working Paper7552.

[3] Hitt, M.A., R.E. Hoskisson, and H. Kim, (1997), "International diversification: Effects on innovation and firm performance in product-diversified firms, Academy of Management Journal, 40(4),767-798.

[4] Kogan,Leonid,D,Papanikolaou,A.Seru,andNoahStoffman,(2017),"Technologi- cal innovation, resource allocation, and growth", Quarterly Journal of Economics, 132(2),665-712.

[5] Koh, Ping-Sheng, and David M. Reeb, (2015), "Missing R&D", Journal ofAccount- ing and Economics, 60(1),73-94.

[6] Kothari, S.P., Ted E. Laguerre, and Andrew J. Leone, (2002), "Capitalization versus Expensing:EvidenceontheUncertaintyofFutureEarningsfromCapitalExpendi- tures versus R&D Outlays", Review of Accounting Studies, 7(4),355-382.

[7] Kafouros, M., C. Wang, E. Mavroudi, J. Hong, and C.S. Katsikeas, (2018), "Ge- ographic dispersion and co-location in global R&D portfolios: Consequences for firm performance", Research Policy, 47(7), 1243-1255.

[8] Lerner, Josh, and Julie Wulf, (2007), "Innovation and Incentives: Evidence from Corporate R&D", Review of Economics and Statistics 89.

[9] Loughran, Tim, and Bill Mcdonald, (2011), "When is a liability not a liability? Tex- tual analysis, dictionaries, and 10-Ks", The Journal of Finance, 66(1), 35-65.

[10] Loughran, Tim, and Bill Mcdonald, (2016), "Textual analysis in Accounting and Finance: A Survey", Journal of Accounting Research 54(4), 1187-1230.

[11] Riitta,Katila,(2000),"Measuringinnovationperformance",InternationalJournalof Business Performance Measurement, Vol 2,180-193.

[12] Tan, J., and M.W. Peng, (2003), "Organisational slack and firm performanceduring economic transitions: two studies from an emerging economy", StrategicManage- ment Journal, 24(3),1249-1263.

[13] Zahra,S.A.,DuaneIreland,andM.A.Hitt,(2000),"Internationalexpansionbynew venturefirms:Internationaldiversity,modeofmarketentry,technologicallearning, and performance", Academy of Management Journal, 43(5),925-950.

[14] Zvi, Griliches, (1990), "Patent statistics as economic indicators: A survey",Journal of Economic Literature, 28(4),1661-1707.

★★★