Identifying Linguistic Innovators in the Business Community

Paul Vicinanza
Graduate School of Business
Stanford University
Stanford, CA 94305
pvicinan@stanford.edu
Mentor: Michael Hahn

Abstract

In conventional management wisdom, innovation defines modern organizational success. To compete in an increasingly global economy, firms must produce at the technological frontier and modernize their managerial practices. This project asks whether the organizational characteristic of innovativeness can be quantified through the discourse executives use in discussing their firm. To accomplish this task, I train an LSTM to minimize sentence perplexity on a data set of quarterly earnings calls by publicly traded firms. I develop a novel method to measure linguistic innovativeness by exploiting inter-year heterogeneity in sentence-level perplexity and use it to identify innovative firms. I segment linguistic innovativeness into two separate variables, median sentence innovativeness and tail sentence innovativeness. Median innovativeness quantifies the extent to which the firm's discourse conforms to the standard business lexicon while tail innovativeness quantifies the "boundary pushing" innovations of the firm. I find that both median and tail innovativeness negatively correlate with stock performance. However, if a firm is high in both median and tail innovativeness, the negative impacts of innovativeness stock returns disappears. These effect are substantial in magnitude. The returns to a firm with high median innovativeness are 3.5 percentage points lower per year. This finding questions traditional scholarly accounts of innovation which consider it unambiguously positive.

1 Introduction & Prior Research

Despite driving firm performance and global economic growth, scholars have little intuition behind the determinants of innovation in firms. Most accounts treat the level of innovativeness as an attribute of the firm. More innovative firms perform better in the marketplace and crowd-out static firms [1]. For many industries, a great tension lies between exploring new frontiers and exploiting old ones [2]. Pharmacology, for example, relies both on the discovery of novel compounds and heavily marketing old ones before their exclusivity period end.

Yet the origins of innovation in firms remains unsettled in the literature. The dynamic capabilities literature posits that the ability to innovate and adapt lies in concrete routines and processes of the firm [3]. Alternative viewpoints highlight the role executive leadership in guiding firm-level innovation. Popular narratives of visionary CEOs (e.g. Steve Jobs) support this account, although empirical data has been mixed [4, 5]. Although these theories are not mutually exclusive, they highlight difficulties both with identifying and measuring the innovativeness of the firm and understanding its downstream consequences.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

For organizational scholars, this project offers new tools to quantify attributes of firms through their linguistic characteristics. NLP deep learning enables the extraction of abstract concepts from unstructured text data. These techniques are particularly well suited to the large corpora of public corporations such as 10-K earnings statements and the quarterly earnings calls studied in this project. Although this project seeks to quantify innovation, similar projects could examine shareholder sentiment or competitive pressures.

This project should interest the deep learning community for two reasons. First, I demonstrate significant reductions in model perplexity by incorporating sentence level meta-data regarding the origins of the sentences. By training firm industry and year embeddings, and passing this data as part of the input to the LSTM, I reduce model perplexity by approximately 2.5%. Second, I offer a novel application of deep learning frameworks by exploiting intra-sentence heterogeneity by modifying the meta-parameters of the sentence.

2 Modelling Approach

2.1 Defining Linguistic Innovativeness

I use linguistic innovativeness to refer to the property of a segment of speech as novel or unconventional which subsequently becomes conventional over time. The discussion of data analytics and deep learning in the business lexicon would be unprecedented in the 1980s but appears common-place now. Consider linguistic innovativeness as the usage of language which appears uncommon in the time period it's used but common in some future period. Importantly, this definition of innovation does not distinguish between innovations which originate from the focal organization or whether the firm is an early adopter of the lingustic innovation.

2.2 Modelling Strategy

Operationalizing this definition requires a model which calculates the probability of an utterance as a function of the spoken year. Standard deep learning NLP language models model sentences by minimizing sentence perplexity, the inverse likelihood of the model generating the sentence. A statement with a high perplexity in the focal year and a lower perplexity in a future period would be classified as an innovative sentence.

To further quantify, assume the language model calculates the probability of a sentence conditional on its spoken year normalized for the number of words: $PP(s \mid y) = PP(x_1, x_2, \dots x_n \mid y)$. Assume further we segment the years into two sets, early years, y_{early}, and late years, y_{late}. Using this framework, I calculate innovativeness, I_s as the percentage decrease in average sentence perplexity between the y_{early} and y_{late}:

$$PP(s \mid y_{early}) = \frac{1}{\# y_{early}} \sum_{y \in y_{early}} PP(y \mid s)$$

$$PP(s \mid y_{late}) = \frac{1}{\# y_{early}} \sum_{y \in y_{late}} PP(y \mid s)$$

$$I_s = \frac{PP(s \mid y_{early}) - PP(s \mid y_{late})}{PP(s \mid y_{early})}$$

Several possibilities exist to aggregate from the sentence level to the document or firm level, which explicate in Results.

2.3 Model Framework

To calculate sentence perplexity, I implement a single layer LSTM aiming to minimize cross-entropy loss between the model's predicted next word in the sequence and the actual next word in the sequence. A single pass through the model constructs the embedding vector, passes the embedding vector through a dropout layer to the LSTM, runs the output of the LSTM through another dropout layer, and passes the output through a dense layer to obtain the logits needed to predict the next word. Figure 1 provides a visualization of the network architecture.

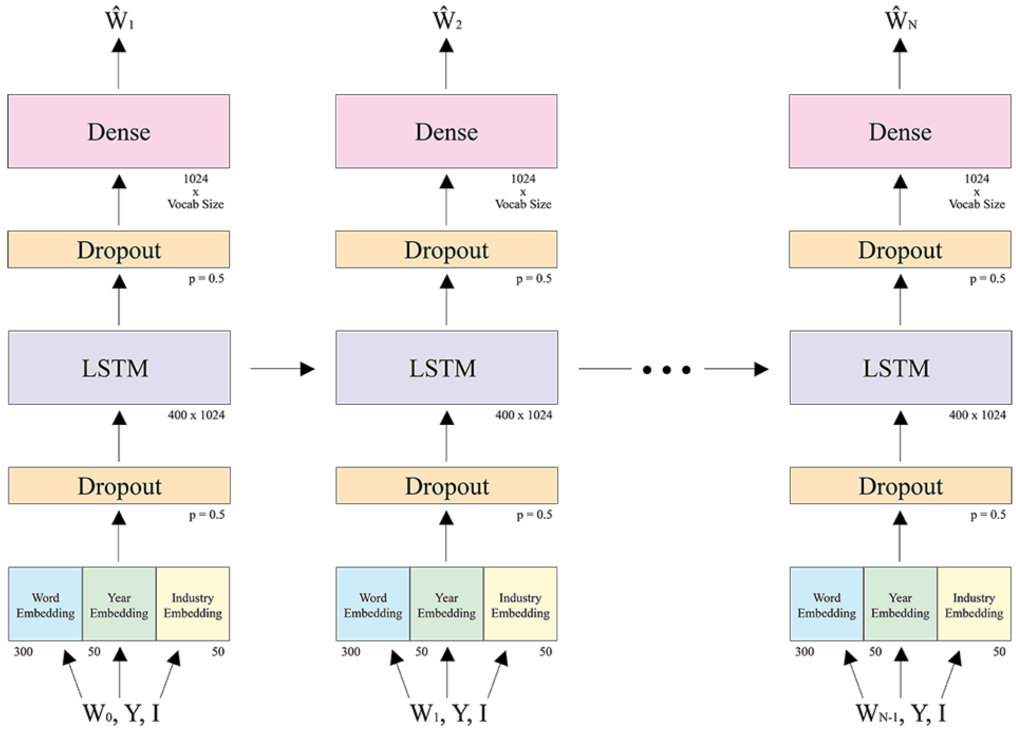

Figure 1: Network Architecture

To model the yearly variations in sentence perplexity and control for the shifting composition of firms, I also train industry and year embeddings in the model. Each statement has an associated industry (from the firm) and year. Industries are defined using the highest level NAICS code. I concatenate the firm and year embeddings for each word and pass the concatenated embeddings as input for the LSTM. To generate the distribution of perplexities for a statement by year, I pass the same statement and industry while modifying each year.[1] The baseline model is the model without yearly or industry embeddings. If the inclusion of these embeddings does not reduce sentence perplexity, then further modelling of linguistic innovation would be moot. Table 1 reports the model parameters and differences between the base model and the final model.

Table 1: Model Parameters

Layer	Base Model	Final Model
Word Embeddings	Vocab Size x 300	Vocab Size x 300
Industry Embeddings	None	Num Industries x 50
Year Embeddings	None	Num Years x 50
LSTM	300 x 1,024	400 x 1,024
Dense Layer	1,024 x Vocab Size	1,024 x Vocab Size

3 Data

I have obtained over 100,000 scrapped and parsed quarterly earnings calls (QECs) of publicly traded companies from the crowd-sourced financial services website seekingAlpha, totaling over half a billion tokens. In these calls, analysts ask questions regarding company performance and executives respond with answers in a back-and-forth Q&A flow. Statements have been parsed by speaker such

[1]Special thanks to Michael, my project mentor, for this suggestion

that I know which analyst/executive is speaking. I have filtered statements to only include utterances by executives in the firms and this comprises my data sample.

Unique to this data is that documents are nested within firms over time. Analysis of temporal word embeddings found that words such as "deep", "learning", "analytics", "cloud" etc. experienced meaningful shifts in the embedding space over time. In 2006, the nearest neighbors of "learning" were "schools" and "children" while the nearest neighbors of "learning" in 2016 were "data", "big", etc., suggesting that the term "learning" experienced semantic change in this time period. Presumably the change was not instantaneous and there are some firms who began using "learning" in conjunction with data science prior to its adoption by other firms.

3.1 Data Processing

Statements were parsed into sentences and tokenizing using Python's NLTK package. The only additional data cleaning / processing step performed was to convert all characters to lowercase. A 300-dimensional skip-gram negative-sampled word embedding model was trained on the entire corpus with a window size of 10 and a negative sampling count of 5.[6]

After I finished training the embedding, I removed sentences longer than 50 tokens (after appending start and end tokens) to improve LSTM computational efficiency. The resulting data had roughly 11 million sentences totalling 350,000 unique tokens. Building the vocabulary with a minimum term frequency of 100 occurrences resulted in an approximate vocabulary size of 20,000 words. Data was split 99% train 1% dev.

Firms were fuzzy matched on firm name to gvkey to obtain industry category and stock price from Compustat, a database of financial and market information on publicly traded firms. Industry category is defined by the 2017 firm classification according to the North American Industry Classification System (NAICS) resulting in 19 unique industries. An oversight on my part, the year of the call was not cleaned prior to model training, and some calls with dates in the early 1900s or 2100s were included in training. Thus there were 17 years in the year embedding matrix. For all functional purposes the data spans from 2006 to 2016. I focus on the years between 2008 and 2016 for the subsequent analysis.

4 Model Performance

4.1 Baseline Model Comparison

To first establish a baseline, I trained a model without industry and year embeddings. The LSTM receives and 300-dimensional input instead of a 400-dimensional input but there are not other changes. Figure 2 reports the model results for both train and dev. The y-axis reports the per-word average loss and the x-axis reports the number of batches through the data. Both models are trained using a learning rate of 0.001, a batch size of 256, and trained for two epochs (24 hours). The base model outperformed the more complicated model substantially in early periods of training. Given the addition of over 100,000 parameters in the more complicated model this is not surprising. Interestingly, the base model outperformed the final model throughout training on the train data (although the difference is minute). Performance on the development set surpasses train set performance due to multiple dropout layers. With regards to dev set performance, the final model surpassed the base model after 4 million training examples. Final dev perplexity for the base model was 32.25 and the final dev perplexity for the industry and year embeddings model reached 31.56. This approximates to a 2.2% reduction in model perplexity. Given the only additional information was the industry and year embeddings, they added pertinent information to the language model.

4.2 Exploring Year Embeddings

If the model correctly learned the year embeddings, we would expect that the year which minimizes the perplexity of a sentence would be the actual year of the sentence (e.g. if the sentence came from a 2009 call transcript, we would expect the lowest perplexity of the sentence when passing in 2009 as the input year).

I test this by passing each sentence through the data 9 times, once for each year between 2008 and 2016. Figure 3 reports the results of this analysis. Each line represents the average perplexity for all

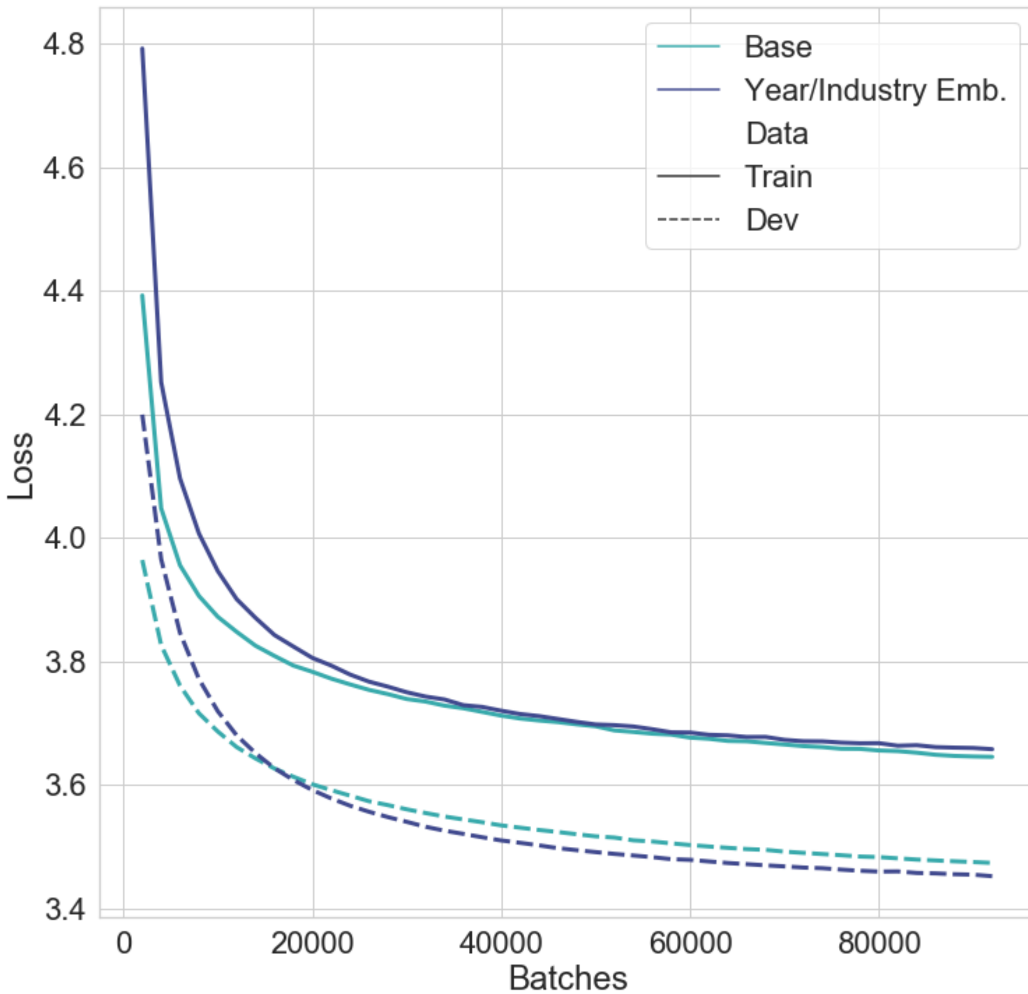

Figure 2: Model Performance

sentences from a particular year. The lightest line represents the average perplexities for all sentences from 2008. The perplexities are normalized by subtracting the mean perplexity from 2008. This normalization enables analyzing each year by its trend over time, rather than its magnitude. The x-axis indicates the "fake" year passed the model. For the line indicating sentence year 2010, the value at simulated year 2016 is the average difference in model perplexity for all sentences passed to the model with the year 2008 and the year 2016.

Results of this analysis strongly support inference using these year embeddings. The lowest perplexities for years 2008, 2010, 2011, 2012, 2013, 2014, and 2015 are the actual years. The lowest perplexities for years 2009 and 2016 are 2008 and 2015 respectively. Furthermore, the graphs show clear trends in perplexity over time. For sentences from 2008, perplexity steadily increases while it steadily decreases for sentences from years 2014, 2015, and 2016.

5 Predicting Stock Market Returns

With model sanity established, I implement the procedure outlined in Modelling Strategy. I exclude 2006, 2007, and 2008 from the analysis to avoid bias from the Great Recession. I define the early years as 2009, 2010, and 2011 and the late years as 2014, 2015, and 2016. I calculate the innovativeness of all sentences from originating from 2009 to 2011, corresponding to the early years of the model. A positive innovativeness score indicates that the sentence has lower perplexity in the future than it does

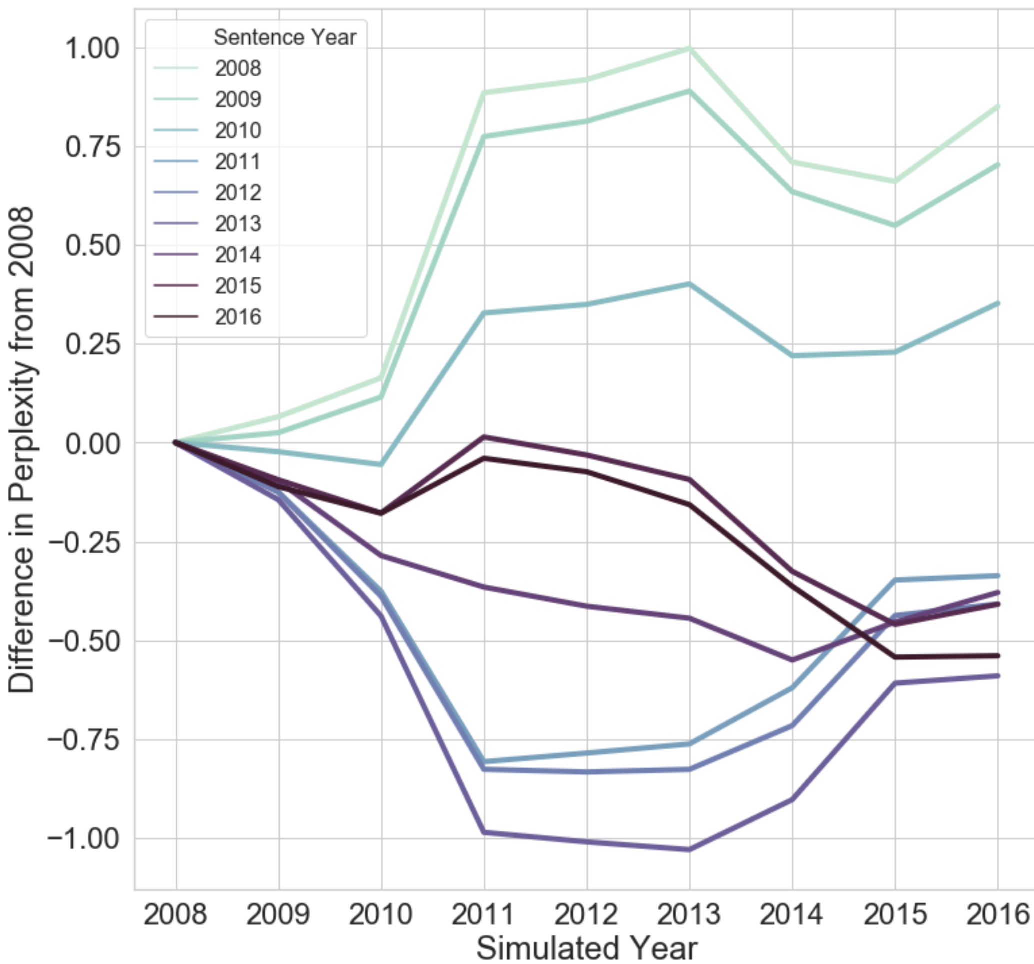

Figure 3: Simulated Perplexities

in its originating period. Mean innovativeness, unsurprisingly, is negative. I aggregate this measure to the firm-level using several aggregation strategies. *Median innovativeness* measures the firm's median sentence-level linguistic innovation score and *tail innovativeness* is the 99th percentile sentence-level innovation score. I also construct indicator variables to ease interpretation of the results. *Median innovativeness dummy* takes a value of 1 if the firm's *median innovativeness* is above firm population average and 0 otherwise. *Tail innovativeness dummy* is constructed the same way.

Table 2 reports a selected list of prominent firms ordered by median innovativeness. The fact that LinkedIn and Amazon are the two most innovative firms, while several traditional brick-and-mortar retailers who are struggling with the E-commerce revolution are the least innovative, provides good credence for the measure. Several other interesting trends are apparent in the data. Firstly, there exist clear clusters of firms by industry. Pharmaceuticals appear near the top of the list while automobile manufacturers cluster near the bottom. Paired firms appear proximate, Visa and Mastercard, HP and IBM, Coca Cola and Pepsi, Nike and Under Armor, etc. suggesting an accurate measure. The fact that AT&T is near the bottom of the list shouldn't surprise anyone who has relied upon their cellular network. Several unexpected findings include Google being in the middle of the list and Best Buy occupying a top spot. Most Silicon Valley firms did not make the analysis as they would need to go public prior to 2012.

I use the company's five-year stock returns beginning in January 2012 as the dependent variable in models 1 to 3. Average returns in this period approximate an 86 percentage increase in firm valuation using the compound returns formula. In model 4, I construct a new binary dependent variable, *hit*

Table 2: Firm Linguistic Innovativeness

Firm	**Median Innovativeness**	**99th Percentile Innovativeness**	**Business**
LinkedIn	74	2.1	Professional social network
Amazon	73	2.0	E-commerce
Best Buy	69	1.7	Retail
Monsanto	68	1.4	Agrochemical
Novartis	67	3.1	Healthcare
GlaxoSmithKline	66	1.4	Pharmaceuticals
Bristol-Meyers Squibb	66	3.9	Pharmaceuticals
Salesforce	64	4.3	Software
CVS	64	1.3	Retail pharmacy
Phizer	63	3.1	Pharmaceuticals
Under Armor	63	1.5	Clothing
Coca Cola	60	1.6	Beverage
Wells Fargo	60	1.0	Banking
Nike	58	2.0	Clothing
Shell	58	2.6	Petroleum
NVIDIA	58	2.2	Computing
Pepsi	57	1.1	Beverage
Siemens	57	1.3	Conglomerate
Google	56	1.4	Technology
Comcast	56	1.2	Telecommunications
Visa	55	3.1	Financial Services
Procter and Gamble	54	0.4	Consumer goods
Netflix	53	1.8	Movie Rentals
Bayer	53	1.2	Pharmaceuticals
Lockheed Martin	52	1.2	Aerospace
AMD	52	2.7	Semiconductors
Cisco	52	1.5	Computing
Mastercard	51	1.7	Financial Services
Hewlett Packard	51	1.4	Computing
Verizon	50	1.0	Telecommunications
IBM	49	1.2	Computing
BP	49	1.4	Petroleum
Yahoo	49	1.1	Technology
General Motors	48	1.5	Automobile
Toyota	48	0.0	Automobile
Fiat	47	0.6	Automobile
Walt Disney	46	0.2	Entertainment
Ford Motor Company	46	2.0	Automobile
The Boeing Company	45	0.6	Aerospace
JPMorgan Chase	45	1.3	Banking
Hasbro	43	1.1	Consumer goods
Halliburton	42	0.0	Petroleum
AT&T	42	0.7	Telecommunications
Goldman Sachs	41	0.4	Banking
Canon	38	6.9	Optical devices
Dick's Sporting Goods	33	0.2	Retail
Barnes & Noble	32	0.0	Retail
OfficeMax	21	0.0	Retail

stock, which takes a value of 1 if investing in the stock in 2012 would have tripled returns by 2016. No controls are included in the models at this point in time. Models are estimated using linear regression with industry fixed effects and industry cluster-robust standard errors.

Table 3: Main Results - Linear Regression Analysis

	(1) Returns	(2) Returns	(3) Returns	(4) Hit Stock
Median Innovativeness	-0.613*	-0.919**		
	(-2.15)	(-3.46)		
Tail Innovativeness	2.225	-27.42**		
	(0.76)	(-3.23)		
Median $\times$ Tail Innov.		50.83**		
		(3.83)		
Median Innovativeness Dummy			-0.200***	-0.0138
			(-4.56)	(-1.93)
Tail Innovativeness Dummy			-0.206**	-0.0404**
			(-2.89)	(-3.82)
Median $\times$ Tail Innov. Dummy			0.339***	0.0482***
			(4.96)	(4.13)
Constant	1.328***	1.495***	1.141***	0.0743***
	(8.67)	(10.97)	(33.61)	(15.49)
R^2	0.002	0.004	0.003	0.002
Observations	1862	1862	1862	1862

t statistics in parentheses
Industry Fixed Effects and Clustered Robust SEs
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5.1 Results

Table 3 reports the results of this analysis. At first pass, median linguistic innovation is weakly negatively correlated with future stock returns and tail linguistic innovation is not significant (Model 1). However, interacting the two terms returns a surprising result (Model 2). Both median and tail linguistic innovativeness have strong negative correlations with stock returns, however, the interaction between the two variables is highly significant and positive. Restructuring the variables as dummies permits a simpler interpretation of the results (Model 3). Having above average median or tail innovativeness corresponds to a 20 percentage point decrease in stock returns. However, having both high median *and* high tail innovativeness mostly eliminates the detrimental effects. When examining the probability of becoming a “hit" stock, we observe that high tail innovativeness reduces the probability by 4% (Model 4). Once again, having both high median and tail innovativeness offset the negative returns.

5.2 Discussion

This result runs contrary to my prediction - I predicted that high linguistic innovativeness would correlate with higher stock returns. Instead, I find that linguistic innovativeness has a strong, negative correlation with stock returns. The effect size is large; above average median or tail innovativeness corresponds to 3.5 percentage point lower annual stock returns. However, my findings suggest that the detrimental effect concentrates in firms which fail to go "all in." Firms which have both high median innovativeness and high tail innovativeness do not experience significantly lower returns.

This finding sheds light on the difficulties of non-innovative firms face when embracing the technological frontier. Many firms attempt innovation without truly being innovative firms. Currently every major automobile manufacturer is investing in self-driving car technologies yet lack the routines, organizational structures, and employee base to efficiently implement it. At the same time, they face substantial competitive pressures from Google's Waymo, Tesla, and other Silicon Valley firms better equipped to conquer the new frontier. Alphabet, Apple, Tesla, etc. represent the new generation of innovators; firms high in both median and tail innovativeness. My results suggest that these traditional car companies may be best served staying in the lane. Leave the innovation to the innovators.

6 Conclusion & Future Directions

This study demonstrates a novel application of NLP deep learning technology to answer social science questions. Traditional deep learning asks, conditional on the sentence, can I predict some characteristic (translation, sentiment, rating, etc.)? My methodology inverts this relationship by asking, conditional on the prediction, can I model the sentence? I find that the inclusion of industry and year embeddings in model training significantly decreases model perplexity, suggesting greater appreciation should be given to sentence context when language modelling. Future work could look to pass author gender or nationality - instead of year or industry - to replicate important socio-cognitive differences in human speech.

Stock returns, while useful, are an imperfect benchmark to determine how linguistic innovativeness relates to the true innovative capabilities of the firm. As as next step, I intend to see how this measure predicts firm patenting rates. If my construct works, we would expect to see firms which are high in linguistic innovativeness are also high in patenting output. Additionally, I have little concrete intuition for what median innovativeness and tail innovativeness conceptually measure. Future work will explore what these metrics capture to better understand these stark statistical patterns.

7 Code

I received the data scrapped and parsed, but did all data cleaning and preparation myself. I wrote all the code myself with several exceptions. Initial network structure was taken from (https://github.com/ap229997/LanguageModel-using-Attention/blob/master/model/net.py) but has been modified so substantially it bears no resemblance and the training and testing of the model was adopted from code from assignment 3.

References

[1] Cefis, Elena, and Orietta Marsili. 2006. "Survivor: The Role of innovation in Firms' Survival. *Research Policy*, 35(5): 626-641.

[2] Uotila, Juha, Markku Maula, Thomas Keil, and Shaker A. Zahra. 2009. "Exploration, Exploitation, and Financial Performance: Analysis of S&P 500 Corporations." *Strategic Management Journal*, 30: 221-231.

[3] Eisenhardt, Kathleen M., and Jeffrey A. Martin. 2000. "Dynamic Capabilities: What Are They?." *Strategic Management Journal*, 21(10-11): 1105-1121.

[4] Balkin, David B., Gideon D. Markman, and Luis R. Gomez-Mejia. 2000. "Is CEO Pay in High-Technology Firms Related to Innovation?." *Academy of Management Journal*, 43(6): 1118-1129.

[5] Denti, Leif, and Sven Hemlin. 2001. "Leadership and Innovation in Organizations: A Systematic Review of Factors that Mediate or Moderate the Relationship." *International Journal of Innovation Management*, 16(3): 1240007.

[6] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation Word Representations in Vector Space." *Proceedings of the International Conference on Learning Representations*.

[7] Jozefowicz, Rafal, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. "Exploring the Limits of Language Modeling." arXiv preprint arXiv:1602.02410.