What Are You Saying? Using topic to Detect Financial Misreporting

Brown N C, Crowley R M, Elliott W B.

Journal of Accounting Research, 2020

叶鑫　2021/10/24

Background

- Detection models have long focused on quantitative financial statement as predictive factors (Beneish [1997], Dechow et al. [2011], Bao et al. [2020]).
  - But misreporting firms often manipulate performance metrics and accounting transactions to escape being found fraudulent.
  - Quantitative financial variables are typically backward-looking and less efficient in predicting misreporting.

- Recent studies analyze the textual and linguistic features of management disclosures.
  - But commonly used textual style features (tone, verbal complexity, readability) are difficult to classify as deceptive, because disclosure narratives can be influenced by individuals' expectations and motivations (Douglas and Sutton [2003]).

Motivation

- Latent Dirichlet Allocation technique (LDA) is an unsupervised and unstructured probabilistic model that learns the latent thematic structure of words within a corpus of documents.

- LDA is widely used in practice by Internet search engines to guide keyword selection and improve correlations between search terms and web content.

- Bao and Datta [2014] examined how the types of risks discussed in 10-K filings influences investors' risk perceptions. But LDA is relatively new to accounting fraud detection.

- Therefore, this article uses LDA to detect and quantifie the thematic content (topic) of the whole annual report narratives, focusing on what is being disclosed by management rather than how.

Research question

- Do disclosure topics improve the detection of intentional financial misreporting, relative to quantitative financial measures?
  - Regulatory oversight is more difficult for financial statement narratives, leaving more room to obfuscate or mislead.
  - The textual narratives in annual report filings contain forward-looking statements (Van Buskirk [2018]).

- Do disclosure topics improve the detection of intentional financial misreporting, relative to aggregate textual style features?
  - Commonly used textual measures do not reflect the context or meaning of management disclosures, thereby limiting the inferences that can be drawn.

Research Contents

1. We run the LDA algorithm on the 10-K filings using rolling five-year windows. The topics discovered in each five-year window are then used in the subsequent year.

2. We first evaluate the semantic validity of the LDA output and the ability to detect misreporting in samples on an annual basis.

3. Our next analyses assess the usefulness of topic in detecting misreporting out of sample, compared to a comprehensive set of quantitative financial and textual style variables.

4. Robust tests: misreporting over time, firms with repeated misreporting events, an expanded set of financial and textual style measures, topics derived from (MD&A) section.

Research Conclusion

- Using both human and machine-based procedures, we find that LDA produces a coherent set of semantically meaningful topics that capture the economic content of annual report filings.

- We find that topic provides significant incremental predictive power over our collection of F-score and Style variables.

- We also find that topic significantly improves the detection of serious revenue-recognition and core-expense errors.

- Our inferences are robust to a series of additional robust tests.

Related research

- Hoberg and Lewis [2017] use LDA to examine whether misreporting firms provide abnormal MD&A disclosures and the underlying incentives for this behavior.
  - Our study considers the thematic content of the entire 10-K filing.
  - We employ a 5-year-rolling-window estimation procedure that accounts for the time-varying nature of the topics discussed by management.

- Wei Dong and Shaoyi Liao [2018] use SFL to extract financial social media data for corporate fraud detection.
- Patricia and Alisa [2020] use a deep learning method (HAN) for detecting and interpreting financial statement fraud, which is more sophisticated and advanced.
- Jeremy and Edwige [2021] use GBRT (RUSBoost) and hundreds of structured variables to detect fraudulence.

Research Contribution

- First, we extend the literature by documenting that the topics generated by LDA are useful for identifying intentional misreporting, either on a standalone basis or in combination with standard prediction variables.

- Second, We exploit a robust machine learning tool that quantifies what is being disclosed in annual financial reports (as opposed to how).

- Further, our real-time prediction method considers the time-varying and fluid nature of management communications, contrasting with prior work based on word dictionaries, which are fairly static and easily identifiable by firms.

- Lastly, our study has important practical implications for regulators and corporate monitors.

Research Data & Predictive variable

- Textual data: 131,528 annual 10-K filings in 1994-2012. We use the full set of filings to generate topics, as this improves the LDA's convergence.
  - Provide comprehensive coverage of the firm and its activities throughout the fiscal year.
  - Maximize the number of firm-year observations in our prediction tests.
  - Other textual features: readability, complexity, tone et al.

- Financial and market data: we merge the sample of 10-K topic filings with Compustat and CRSP which exclude financial firms and firm-years with missing data.
  - Expand Dechow [2011] F-score model: accrual quality, firm performance, market pressures, firm size, audit quality et al.

Research Data & Predicted variable

- SEC AAERs which covers accounting and auditing violations before 2012, considering the latent period of accounting fraud.
  - We identify 505 misreported 10-K filings for 192 unique firms issued from 1994 to 2010.
- Intentional misreporting events from the Audit Analytics Non-Reliance Restatements database (AA) since 2000.
  - We identify 527 misreported filings issued by 245 unique firms from 2000 to 2012.
- A customized automated search of amended 10-K filings for misreporting.
  - We identify 697 misreported filings across 553 unique firms from 1994 to 2012.

1. Variants of the words "fraud" or "irregularity": "... fraud* ...," "... irregular* ...," "... materially false and misleading ...," "... violat* of federal securities laws ...," "... violat* securities exchange act ..."
2. Presence of related SEC or Department of Justice (DOJ) investigations: "... sec ... investigat* ...," "... investigat* ... sec ...,"

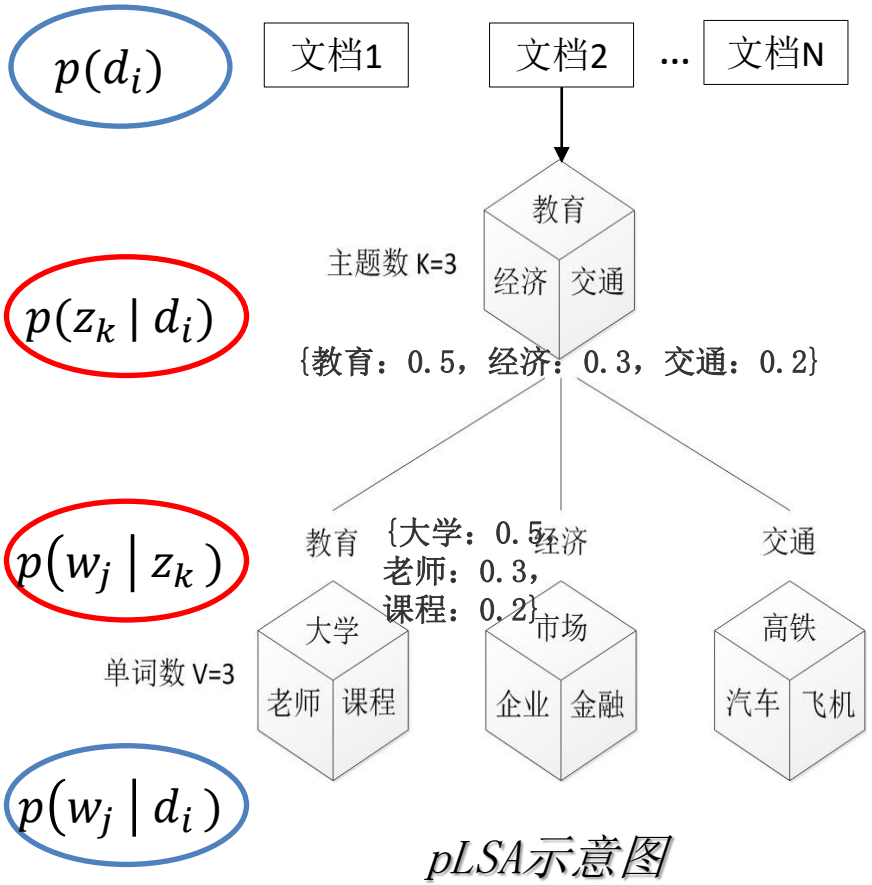
Research Data: Predicted variable

Distribution of Financial Misreporting

	Panel A: AAERs			Panel B: AA Irregularities			Panel C: 10-K/A Irregularities		
Year	Observations	Frequency	Percentage	Observations	Frequency	Percentage	Observations	Frequency	Percentage
1994	786	0	0.00				786	2	0.25
1995	1,043	6	0.58				1,043	2	0.19
1996	1,634	17	1.04				1,634	17	1.04
1997	2,250	23	1.02				2,250	16	0.71
1998	2,308	40	1.73				2,308	12	0.52
1999	2,195	47	2.14				2,195	13	0.59
2000	2,041	51	2.50	2,041	36	1.76	2,041	21	1.03
2001	2,021	44	2.18	2,021	39	1.93	2,021	18	0.89
2002	2,391	51	2.13	2,391	60	2.51	2,391	27	1.13
2003	2,936	60	2.04	2,936	81	2.76	2,936	53	1.81
2004	2,843	52	1.83	2,843	77	2.71	2,843	70	2.46
2005	2,678	39	1.46	2,678	75	2.80	2,678	65	2.43
2006	2,608	19	0.73	2,608	39	1.50	2,608	78	2.99
2007	2,549	18	0.71	2,549	19	0.75	2,549	53	2.08
2008	2,535	13	0.51	2,535	14	0.55	2,535	48	1.89
2009	2,564	15	0.59	2,564	28	1.09	2,564	65	2.54
2010	2,424	10	0.41	2,424	24	0.99	2,424	48	1.98
2011				2,330	19	0.82	2,330	44	1.89
2012				2,178	16	0.73	2,178	45	2.07

	Panel A: AAERs			Panel B: AA Irregularities			Panel C: 10-K/A Irregularities		
Year	Observations	Frequency	Percentage	Observations	Frequency	Percentage	Observations	Frequency	Percentage
All firm-years	37,806	505	1.34	32,098	527	1.64	42,314	697	1.65
No. of firms	6,423	192		5,082	245		6,588	553	
Prediction firm-years	29,785	419	1.41	19,866	234	1.18	34,293	648	1.89
No. of firms	5,259	162		3,916	148		5,427	513	

Research Method: LDA Model



pLSA示意图

w_j在文档 d_i 中出现的次数除以 d_i 中词语总数目

- 生成文档的整个过程便是选定要写的文档d_i生成主题z_k，确定主题生成词w_j，重复多次得到文档d_i。
- pLSA就是根据文档反推其主题分布
- 在这个过程中，我们并未关注词和词之间的出现顺序，即pLSA是一种词袋方法。

$$P(w_j|d_i) = \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i)$$

故得到文档中每个词的生成概率为：

$$P(d_i, w_j) = P(d_i)P(w_j|d_i)$$

$$= P(d_i)\sum_{k-1}^{K} P(w_j|z_k)P(z_k|d_i)$$ 求MAX值

- LDA在PLSA的基础上，为主题分布$p(z_k \mid d_i)$和词分布$p(w_j \mid z_k)$分别加了两个Dirichlet先验分布，将其随机化，通过这些分布生成的观测值（即实际文本）来反推最优分布情况.

Research Method: LDA Model

- In short, LDA is a probabilistic process that condenses the vocabulary in a collection of documents into a dictionary of topics and a set of topic weights.

- Important Input: number of topics, BOW word encoding based on 10-K filings
- Model: gensim.models. ldamodel(text, num_topics=31, id2word = dictionary, passes=20)
- Output: $p(z_k \mid d_i)$, $p(w_j \mid z_k)$, and we need to label the topic manually.

- A unique advantage of LDA is that it does not require predetermined word dictionaries or topic categories, which reduces researcher bias, as foreknowledge of document content does not affect the topic classifications.

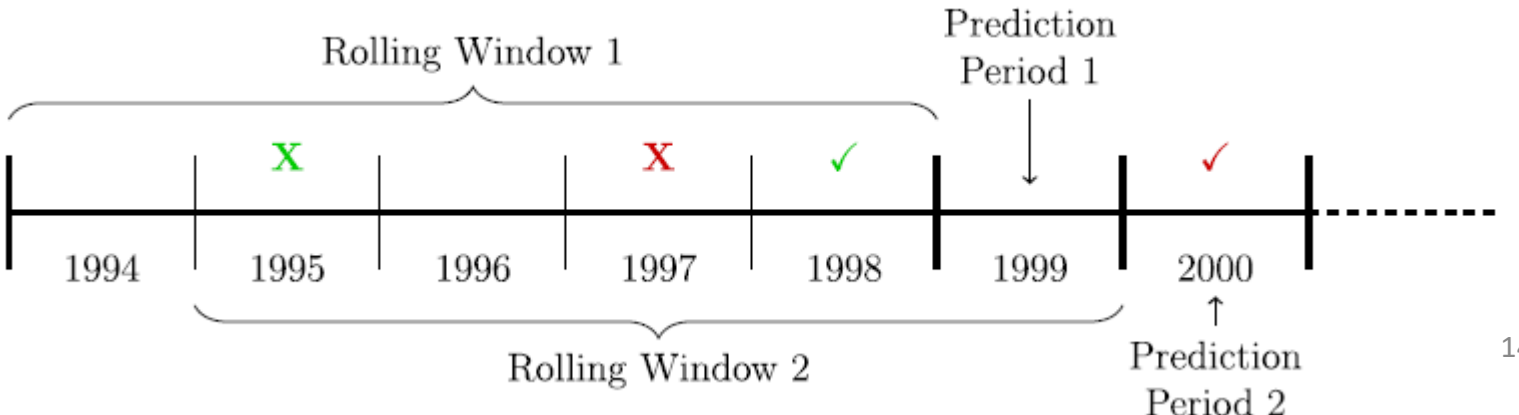
- Furthermore, the algorithm can classify the content of large collections of textual narratives.

Research Method: LDA Model

- We select topics in 14 rolling five-year windows over our sample period, due to macroeconomic or industry trends, changes in disclosure requirements, or managerial turnover.

- We use these 31 topics (optimal in detecting irregularities drawn from amended 10-K filings) and their word weights to compute the prediction period's topic weights and normalize by scaling the sum of the weights of all topics identified in the filing.

- We further orthogonalize the topic proportions to 2-digit SIC to adjust for unobserved industry effects as the final topic.



Empirical result: Evaluation of LDA Topic

- We get 64 combined topics based on the Pearson correlation of the word weights within the annual 31 topics. The best thresholds is 11% from 1% to 90% in 1% intervals.

- To determine the underlying content of each combined topic, we generate a list of the highest weighted phrases and sentences associated with each topic.

- We note that the LDA algorithm performs well in identifying narrative content that relates distinctively to changes in firms' financial performance and their financing activities.

1) Decrease in income compared to prior periods: compared to, gross profit, other income, company contributed, operating income, company expects, gross margin, income decreased, capital expenditures, decreased to Management fee income decreased in 1998 to $0 as compared to $1.4 million in 1997. The Company's gross profit margin decreased to 59% in Fiscal 1996, compared to 65% in Fiscal 1995.
2) Increase in income compared to prior periods: compared with, gross margin, income was, operating income, gross profit, other income, fiscal compared, income taxes, non-interest income, profit was

Empirical result: Evaluation of LDA Topic

- If the set of words from the LDA model is coherent, then the human subjects should easily identify the intruder word at a rate that is significantly higher than random chance. We use 3 of the 10 most probable words of topic1 and 1 intruder word from topic2.
  - The human-subjects task produces an average accuracy rate of 40%(>25%).

- Taken together, our qualitative and quantitative evaluation methods suggest that the LDA algorithm provides a valid set of semantically meaningful topics

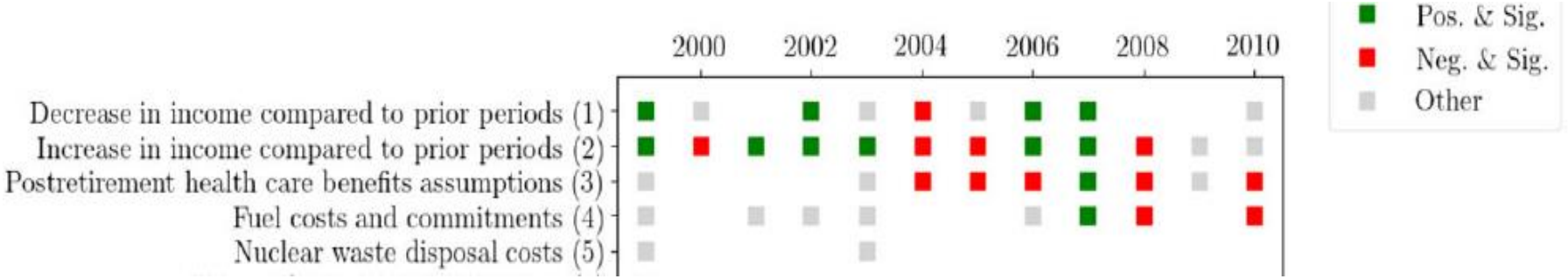
In-Sample detective Value of topic

- We estimate equation for the five-year window preceding each of the out-of-sample prediction years in each of the three misreporting samples (AAER, AA, 10-K/A).

$$\log\left(\frac{misreport_{i,t}}{1-misreport_{i,t}}\right) = \alpha + \sum_{j=1}^{17}\beta_j F\text{-}score_{j,i,t} + \sum_{j=1}^{20}\beta_{j+17} Style_{j,i,t} + \sum_{j=1}^{31}\beta_{j+37} topic_{j,i,t} + \varepsilon_{i,t}, \quad t \in [T-5, T-1]$$

$i \in$ Firms.

- For each prediction year, we present green (red) boxes if at least one subtopic for a given combined topic loads as positive (negative) and significant at the 10% level or greater, and all other subtopics are insignificant.



Topic versus Financial Variables (RQ1)

- The AUCs for the topic model also exceed the 0.50 threshold in all three samples, indicating the ability of thematic content to independently detect various forms of financial misreporting. The predictive value of topic is markedly.
- Both predictors serve as complementary warning signals in the detection of misreporting.

Panel E: AUC statistics (10-K/A irregularities)

Prediction model	AUC
F-score	0.589***
topic	0.616***
topic and *F-score*	0.630***

Panel F: Difference tests (10-K/A irregularities)

		F-score	*topic* and *F-score*
topic	Diff. in AUC	0.027**	−0.014*
	p-value	(0.08)	(0.08)
topic and *F-score*	Diff. in AUC	0.041***	
	p-value	(0.00)	

Topic versus Textual Style (RQ2)

- The benchmarking tests for the AAER and AA samples indicate that topic by itself better predicts misreporting than the standalone Style model.
- We observe that a joint model of topic and Style outperforms the individual Style model by almost 6% in the AAER sample

Panel C: AUC statistics (AA irregularities)

Prediction model	AUC
Style	0.581**
topic	0.616**
topic and *Style*	0.614**

Panel D: Pooled ROC AUC difference tests (AA irregularities)

		Style	*topic* and *Style*
topic	Diff. in AUC	0.035*	0.002
	p-value	(0.06)	(0.85)
topic and *Style*	Diff. in AUC	0.033	
	p-value	(0.13)	

Joint Predictive Value

We find that the three-vector model performs well in detecting misreporting out of sample. The AUCs across the three samples are well.

Panel A: AUC statistics (AAERs)	
Prediction model	AUC
F-score and *Style*	0.719***
topic, *F-score*, and *Style*	0.752***
topic	0.680***
topic and *F-score*	0.742***
Panel C: AUC statistics (AA irregularities)	
Prediction model	AUC
F-score and *Style*	0.606***
topic, *F-score*, and *Style*	0.635***
topic	0.610***
topic and *F-score*	0.632***
Panel E: AUC statistics (10-K/A irregularities>)	
Prediction model	AUC
F-score and *Style*	0.667***
topic, *F-score*, and *Style*	0.670***
topic	0.616***
topic and *Style*	0.669***

The Economic Significance of Topic

- The NDCG@k measure evaluates the ranking quality of each prediction model and ranges from 0 to 1, with higher values indicating greater classification performance, where k is the top 1% percentile of the predicted probability scores.

- To quantify the economic value of topic, we note that the classification rate at the 95th percentile improves by 59% when topic is added to the benchmark F-score and Style model.

Panel A: Classification of AAERs

	50th Percentile		90th Percentile		95th Percentile		NDCG@k
Prediction model	%	Count	%	Count	%	Count	99th
topic	70.29	292	22.75	91	10.39	47	0.113
F-score	72.51	319	22.95	111	11.24	65	0.146
Style	65.73	273	14.60	62	6.78	35	0.079
topic and *F-score*	78.56	332	28.93	140	18.01	86	0.176
topic and *Style*	74.79	312	21.40	91	13.38	52	0.118
F-score and *Style*	76.42	329	24.28	117	14.11	70	0.163
topic, *F-score*, and *Style*	79.43	339	32.90	150	22.44	96	0.188

Controlling for "Repeat Offenders"

- There is a concern that our topic measure could be biased toward identifying repeat offenders or certain types of firms, rather than detecting variations in thematic content when firms misreport.
- Specifically, we remove misreporting firms from the out-of-sample prediction period if misreport is set to 1 in any year during the in-sample estimation window.

Out-of-Sample Classification Performance of topic*: Controlling for Repeat Offenders*

Panel A: Classification of AAERs

	50th Percentile		90th Percentile		95th Percentile		NDCG@k
Prediction Model	%	Count	%	Count	%	Count	99th
topic	65.37	92	15.90	21	2.47	7	0.076
F-score	68.41	109	18.83	37	10.04	23	0.141
Style	61.57	94	10.77	17	2.27	6	0.000
topic and *F-score*	67.19	106	29.69	42	17.20	26	0.162
topic and *Style*	68.87	103	11.22	20	3.09	9	0.076
F-score and *Style*	68.61	111	17.93	40	12.78	26	0.162
topic, *F-score*, and *Style*	67.62	108	27.65	44	12.04	24	0.172
topic, *F-score*, and *Style*	79.43	339	32.90	150	22.44	96	0.188

Conclusion

- Using SEC AAERs and irregularities drawn from financial restatements and annual filing amendments, we find that our topic measure provides significant incremental predictive power over commonly used financial statement and textual style measures.

- Specifically, out-of-sample prediction models that incorporate topic outperform models based solely on financial and textual measures. Further, our results reveal that topic is incrementally and economically valuable in detecting above-normal and high-risk misreporting events, improving prediction accuracy by as much as 59% in the case of SEC AAERs and 50% for irregularity restatements.

- Our results are robust to a battery of sensitivity checks, including alternative definitions of topic, an alternative identification of irregularity restatements, time variations in misreporting, and additional financial and textual variables.

Inspiration

- Machine learning in topic using which generated from the whole10-k filings, MD&A, social media text, conference vocal-to-text.

- Interpretation of the textual topic.

- Sample seletion and forward looking bias.