Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection

LYNNETTE PURDA, DAVID SKILLICORN

Contemporary Accounting Research, 2015.9

Lv Mannі

2021.11.21

Contents

Backgrounds & Motivation

- Fraudulent activity has increased in the post-financial crisis years while at the same time resources allocated to fight fraud have been cut.
- As a result, those responsible for detecting fraud, including auditors, regulators and investors, must be judicious in choosing cost-effective tools to help them with this task.

➢ We want to add to the set of fraud-detection tools.

➢ We want to provide the correlation between the various approaches.

Research Problem

- We develop a statistical method for analyzing the language used in the MD&A section of a firm's annual and quarterly reports. Is it a effective predictor of fraud?
- We find the rate of correct classification to be approximately 82%. And we confirm that a temporal measure of deviations from previous language may provide incremental power in identifying fraudulent reports in time.
- What is the correlation between the various approaches?
- Our method consistently performs well among the eight alternatives we examine, and we find that our measure and F-score measures serve as complements to one.

Contribution

- We make several important contributions to the literature:

1. generating a data-derived language tool that appears to be an effective predictor of fraud,
2. conducting a thorough comparison across both quantitative and language-based detection methods,
3. providing the first indication that a temporal measure of deviations from previous language may provide incremental power in identifying fraudulent reports in time series.

Data and Sample

Sample generation:

1. Find AAER between **October 1999 and March 2009**
2. remove firms from the finance industry
3. remove those without coverage in the COMPUSTAT database (240)
4. record the precise **quarters and financial reports** identified by the SEC as being fraudulent
5. download all available 10-Q and 10-K reports of these firms on the EDGAR database between **1994 and 2006** (4,895)
6. extract the MD&A section and keep all text except legal disclaimers

Model Design

Model Steps:

1. Create a table of word frequencies to identify all words appearing in the sample MD&A sections
2. Use Random Forests (3000, 75%, 55) to sort the words in rank order from most to least predictive
3. Use multiple SVMs (200 words) to predict the probability of each report being truthful

Summary statistics

Panel A: Summary statistics on probability of truthful reporting

	Mean	Median	SD
Full sample	0.80	0.88	0.20
Truthful reports	0.87	0.89	0.11
Fraudulent reports	0.56	0.57	0.23
T and Chi-Square stat for test of difference	61.99	1,000	
(*p*-value)	(0.00)	(0.00)	

Panel B: Predicted versus actual status for the full sample and withheld reports with 80 percent as the fraud threshold

Classification results

	All reports			Withheld reports Not used to discover predictive words			
	Predicted				Predicted		
Actual	Non-fraud	Fraud	Total	Actual	Non-fraud	Fraud	Total
Non-fraud	3,119	649	3,768	Non-fraud	786	154	940
Fraud	223	904	1,127	Fraud	54	226	280
Non-fraud	82.78%	17.22%	100%	Non-fraud	83.62%	16.38%	100%
Fraud	19.79%	80.21%	100%	Fraud	19.29%	80.71%	100%
Correct classification	82.19%			Correct classification	82.95%		
ROC area	0.89			ROC area	0.89		

- The text of the MD&A section can provide important indicators of financial misrepresentation.

Comparison of methods within 10-K reports

- Discussion of alternative quantitative methods

Technique	Description	Source
Quantitative detection techniques		
F-score	Ratio of the predicted probability of fraud based on a logit model of firm financial characteristics, over the unconditional expectation of a financial misstatement.	Dechow et al. (2011)
Capacity difference	Difference between the year-over-year change in revenue less the year-over-year change in number of employees.	Brazel et al. (2009)
M&A activity	Dummy variable equal to one if the firm acquires a target or merges with another firm during the fiscal year. The variable is otherwise equal to zero.	Brazel et al. (2009)
Unexplained audit fees	The residual from a regression of characteristics previously shown to be associated with audit fees on the log of audit fees.	Hribar et al. (2010) with modifications proposed by Price et al. (2011)

- Discussion of alternative language-based methods

Technique	Description	Source
Language-based detection techniques		
Deceptive proportion	Frequency of appearance in the MD&A section of words from a list of deceptive words corresponding to the categories of first person singular pronouns, exclusive words, negative-emotion words, and action verbs. The frequency is then divided by the total number of words in the MD&A.	Newman et al. (2003)
Litigious proportion	Frequency of appearance in the MD&A section of words from the Fin-Litigious list. The frequency is then divided by the total number of words in the MD&A.	Loughran and McDonald (2011)
Uncertain proportion	Frequency of appearance in the MD&A section of words from the Fin-Uncertain list. The frequency is then divided by the total number of words in the MD&A.	Loughran and McDonald (2011)
Negative proportion	Frequency of appearance in the MD&A section of words from the Fin-Negative list. The frequency is then divided by the total number of words in the MD&A.	Loughran and McDonald (2011)

- Common and overlapping words

Deceptive Model (86 words)	Fin-Negative (2,350 words)	Fin-Uncertainty (292 words)	Fin-Litigious (871 words)
Panel A: Most frequent words from each word list within the sample			
or	loss	approximately	contracts
loss	restructuring	may	contract
however	losses	could	litigation
but	impairment	may	claims
although	adverse	must	settlement
carrying	decline	possible	legal
action	adversely	depend	regulatory
without	litigation	might	contractual
driven	claims	depends	laws
except	against	uncertain	court
Panel B: Overlapping words from fixed lists and data-generated list			
loss	critical	approximately	contract
or	loss	believes	legal
		may	settlement

- The data identifies to be a predictive word may be difficult for researchers to identify ex ante.

Comparison of methods within 10-K reports

Panel A: Summary statistics for alternative prediction techniques across annual reports

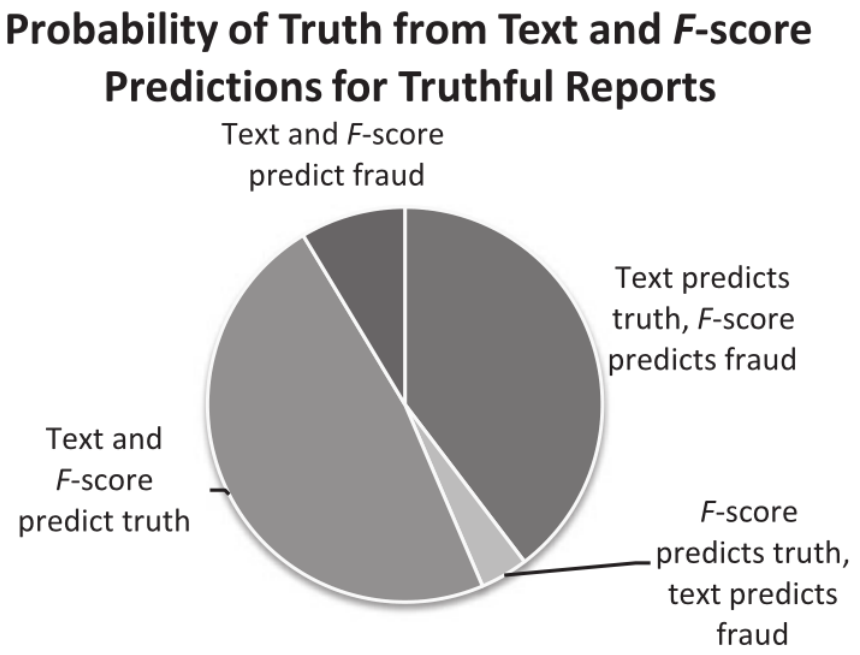
	Probability of truth	*F*-score	Capacity difference	M&A activity	Unexplained audit fees	Deceptive words	Litigious words	Uncertain words	Negative words
Truthful									
Mean	0.877	1.237	0.066	0.017	0.478	1.190%	0.636%	1.314%	1.616%
Median	0.892	0.950	0.049	0.000	0.489	1.130%	0.588%	1.238%	1.551%
N	638	638	608	638	291	638	638	638	638
Fraudulent									
Mean	0.652	1.832	0.090	0.018	−0.208	1.095%	0.502%	1.451%	1.474%
Median	0.704	1.512	0.067	0.000	−0.181	1.041%	0.434%	1.273%	1.358%
N	164	164	157	164	61	164	164	164	164
T-stat for mean diff.	19.38**	−6.16**	−0.85	−0.09	6.06**	2.75**	4.30**	−2.38*	2.22*
(*p*-values)	(0.00)	(0.00)	(0.40)	(0.93)	(0.00)	(0.01)	(0.00)	(0.02)	(0.03)
ROC area	0.87	0.66	0.53	0.50	0.28	0.44	0.40	0.53	0.45

Panel B: Pairwise correlation between alternative prediction methods across annual reports

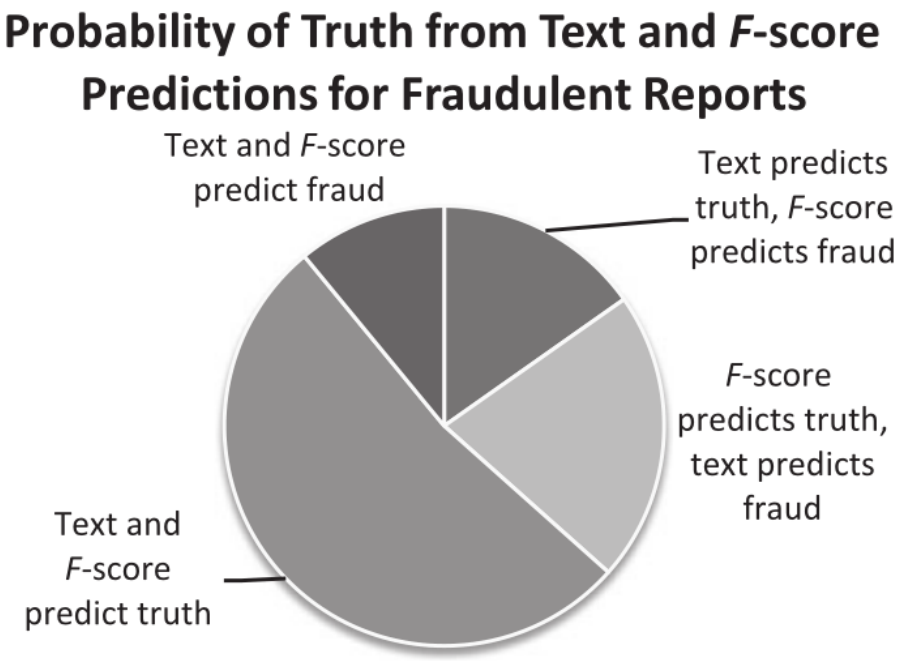
	Probability of truth	*F*-score	Capacity difference	M&A activity	Unexplained audit fees	Deceptive words	Litigious words	Uncertain words	Negative words
Probability of truth	−1.00								
F-score	−0.246**	1.00							
Capacity difference	−0.021	0.008	1.00						
M&A activity	−0.068	0.069	−0.012	1.00					
Unexplained audit	0.301**	−0.094	−0.055	−0.033	1.00				
Deceptive words	0.113**	−0.020	0.004	−0.061	0.033	1.00			
Litigious words	0.223**	−0.149**	0.051	−0.069	0.262**	0.094**	1.00		
Uncertain words	−0.091**	−0.007	0.028	−0.007	0.009	0.158**	0.164**	1.00	
Negative words	0.144**	−0.210**	−0.004	−0.018	0.192**	0.241**	0.471**	0.364**	1.00

Comparison of methods within 10-K reports

- Truthful reports



- Fraudulent reports



- Only 11 percent of frauds go undetected by either method, indicating that textual analysis in combination with an evaluation of quantitative accounting figures can capture the vast majority of financial misrepresentations.

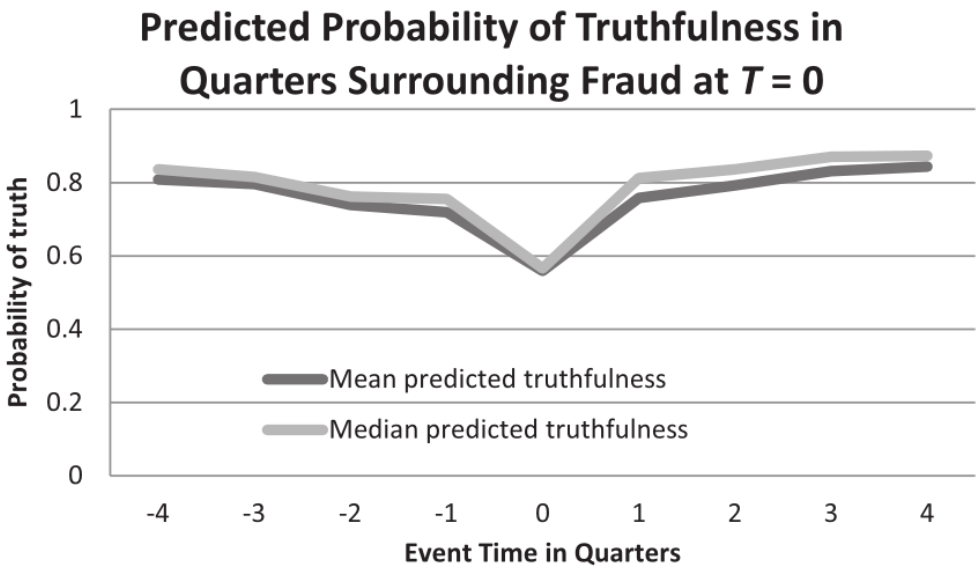
Comparison of methods within 10-K reports

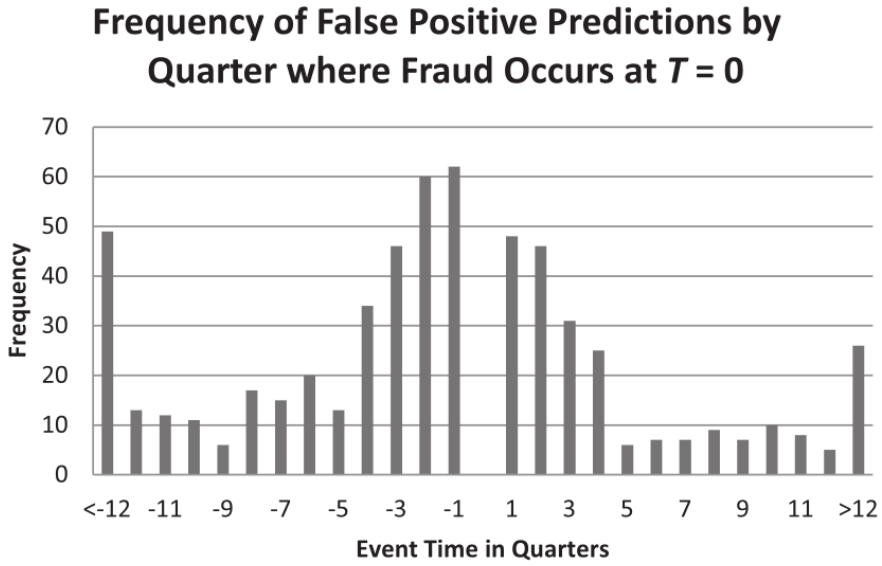
Logit analysis of alternative prediction methods

	Panel A: 802 annual reports (truthful and fraudulent) from firms named in AAER bulletins				**Panel B:** 164 fraudulent annual reports and 164 matches from firms never named in AAER bulletins			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Probability of truth	−8.63**	−12.52**	−8.49**	−8.58**	−8.13**	−9.08**	−7.75**	−8.40**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
F-score	0.20*	0.05		0.23**				
	(0.01)	(0.84)		(0.00)				
Capacity difference	0.31	0.64		0.32	0.89*	0.73		0.83
	(0.48)	(0.49)		(0.45)	(0.03)	(0.20)		(0.05)
M&A activity	−1.23	0.15		−1.16	−2.21*	−1.02		−2.28*
	(0.41)	(0.96)		(0.40)	(0.01)	(0.34)		(0.01)
Unexplained audit		−0.53				0.61		
		(0.01)				(0.06)		
Deceptive proportion			−33.73	−32.68			−65.32	−81.74
			(0.28)	(0.38)			(0.13)	(0.07)
Litigious proportion			−16.80	−24.52			82.37	79.78
			(0.53)	(0.41)			(0.09)	(0.10)
Uncertain proportion			47.95*	47.48*			30.05	29.64
			(0.01)	(0.02)			(0.21)	(0.26)
Negative proportion			6.30	17.03			27.49	30.49
			(0.70)	(0.35)			(0.26)	(0.23)
Constant	5.21**	8.71**	5.13**	4.71**	6.35**	7.33**	5.47**	6.22**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

	Panel A: 802 annual reports (truthful and fraudulent) from firms named in AAER bulletins				**Panel B:** 164 fraudulent annual reports and 164 matches from firms never named in AAER bulletins			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Observations	765	342	802	765	313	122	328	313
Pseudo R^2	0.30	0.39	0.29	0.31	0.27	0.29	0.24	0.29
ROC area	0.87	0.92	0.86	0.87	0.83	0.86	0.81	0.84

Assessing language-based tools in time series





Is the shift of our measure an indicator

	(1) Coefficients (*p*-value)	(2) Coefficients (*p*-value)
Probability of truth	−9.35*	−12.74*
	(0.00)	(0.00)
Change in probability		6.16*
		(0.00)
1-Probability of truth		
Deceptive proportion	5.30	11.94
	(0.62)	(0.36)
Litigious proportion	−37.13	−28.83
	(0.07)	(0.24)
Uncertain proportion	8.71	9.32
	(0.37)	(0.43)
Negative proportion	6.67	0.20
	(0.50)	(0.99)
Constant	5.77*	8.30*
	(0.00)	(0.00)
Observations	4,895	4,449
Pseudo R^2	0.39	0.50
ROC area	0.89	0.93

The statistical approach to textual analysis can be an effective tool for identifying fraud in a quarterly and time-series context.

Conclusion

- We develop a language-based method for detecting fraud using the words in the MD&A sections of annual and interim reports.
- We compare the effectiveness of our method to alternative fraud detection approaches across different samples and find that it consistently performs well. We find F-score is the next most effective indicator, and they serve as complements to one another.
- We examine our language-based tools within a time-series sample and find strong support for our probability-of-truth measure. We find that the change in the probability of truth can provide incremental power in identifying fraud.