Leveraging Financial Social Media Data for Corporate Fraud Detection

Wei Dong, Shaoyi Liao & Zhongju Zhang

Journal of Management Information Systems, 2018.4

Lv Manni

2021.10.10

Contents

- Introduction
  - Background & Motivation
  - Literature Review
  - Research Problem
  - Contribution
- Data and Model Design
- Empirical Results
- Conclusion

Backgrounds & Motivation

- Financial fraud is a serious commercial problem worldwide.
- Existing analytical procedures for corporate fraud investigation highly rely on financial statements.
- However, data in a financial statement are often not appropriate to detect fraud in a timely manner and may contain misleading and fictitious information.
- In recent years, financial social media platforms for investment research have burgeoned. And the opinions and views expressed have been shown to contain value-relevant information and have been used to predict future stock returns.

➢ Can the user-generated content (UGC) on financial social media platforms be useful in assessing the potential risk of corporate fraud?

Literature Review

- **Data Source**- tap into only traditional data sources such as financial statements MD&A and earnings conference calls

- use the MD&A and calls - usually well-planned and prepared in advance

Table 1. Representative Studies of Corporate Fraud Detection and Data Sources

Data type	Indicators	Literature	Data source
Structured data	Numerical financial variables	Cecchini et al. [15]	Financial statements
	Financial ratios	Summers and Sweeney [64]; Dechow et al. [19]; Abbasi et al. [1]	Financial statements
	Nonfinancial variables	Brazel et al. [10]	Financial statements
Unstructured data	Features from language-based textual content	Larcker and Zakolyukina [36]	Earnings conference calls
		Purda and Skillicorn [55]	MD&A section from financial statements
	Features from vocal speech	Hobson et al. [30]	Earnings conference calls
	Social media features	Current study	Financial social media platform, for example, SeekingAlpha

Table 2. Text-based Methods of Corporate Fraud Detection

Technique	Literature	Source of text
Dictionary-based method	Purda and Skillicorn [55]	MD&A section from both annual and quarterly reports
	Larcker and Zakolyukina [36]	Earnings conference calls
	Humpherys et al. [31]	MD&A section of the 10-K report
Statistical method	Cecchini et al. [16]; Glancy and Yadav [26]; Moffitt et al. [47]	MD&A section of the 10-K report
	Goel and Gangolly [27]; Goel et al. [28]	The entire text of the 10-K report

Research Problem

- Is the user-generated content (UGC) on financial social media platforms useful in assessing the potential risk of corporate fraud?
- Is there any incremental value?
- ➢ Based on systemic functional linguistics (SFL) theory, we extracts signals such as sentiment features, emotion features, topic features, lexical features, and social network features, which are then fed into ML classifiers and can detect fraud well.
- ➢ The incremental value really exists.
- Will our model be influenced by rumors in social media?
- ➢ No.

Contribution

- This is one of the first studies to use textual data from social media platforms for corporate fraud detection.
- Use an Systemic Functional Linguistics / SFL-Based Framework for UGC from social media platforms.

Theory and Model Design

- Systemic Functional Linguistics (SFL)

➢ Systemic

➢ Functional : ideational, interpersonal, and textual

概念功能：语言是用来组织、理解和表达我们对世界的看法和我们自己的思想和意识的。

人际功能：语言作为交流的媒介，是创造和维持人际关系的手段。

语篇功能：决定了信息的组织和呈现方式，使得一堆随机排列的句子组成实际的、具有意义的文本。

Theory and Model Design

1. **Ideational Function**: **topics, opinions, and emotions**

- **Topics**

Method: LDA

- **Opinions and emotions**

Method: We use the emotional categories defined in the Linguistic Inquiry and Word Count dictionary to measure sentiment polarity, “assent,” “anxiety,” “anger,” “swear,” and “sadness” emotions.

- **cognitive appraisal: How an individual views a firm’s operations condition.**

Method: Three separate word lists are developed to capture 3 part of cognitive appraisal: (1) overall description of the fraudulent situation; (2) detailed analysis of the fraudulent behavior; and (3) legal judgments and sanctions.

Theory and Model Design

1. **Ideational Function**: Measures of Opinions and Emotion Features

Type	Feature	Measurement
Opinions	Ratio of positive sentiment	Total number of positive words divided by total number of words*
	Ratio of negative sentiment	Total number of negative words divided by total number of words
Emotions	Ratio of assent words	Total number of assent words divided by total number of words
	Ratio of anxiety words	Total number of anxiety words divided by total number of words
	Ratio of anger words	Total number of anger words divided by total number of words
	Ratio of swear words	Total number of swear words divided by total number of words
	Ratio of sadness words	Total number of sadness words divided by total number of words
	Ratio of fraud synonyms words	Total number of synonyms of fraud divided by total number of words
	Ratio of fraud analysis words	Total number of fraud analysis words divided by total number of words
	Ratio of legal judgments words	Total number of legal judgments words divided by total number of words

Theory and Model Design

2. Textual Function

- three information types - writing styles, **genres**, and vernaculars

Genres in a document represent how writers typically use language to respond to recurring situations.

Merkl-Davies and Brennan found that corporate narratives can be regarded as an identifiable genre for business communication with distinctive linguistic properties.

Method: TF-IDF

Theory and Model Design

3. Interpersonal Function

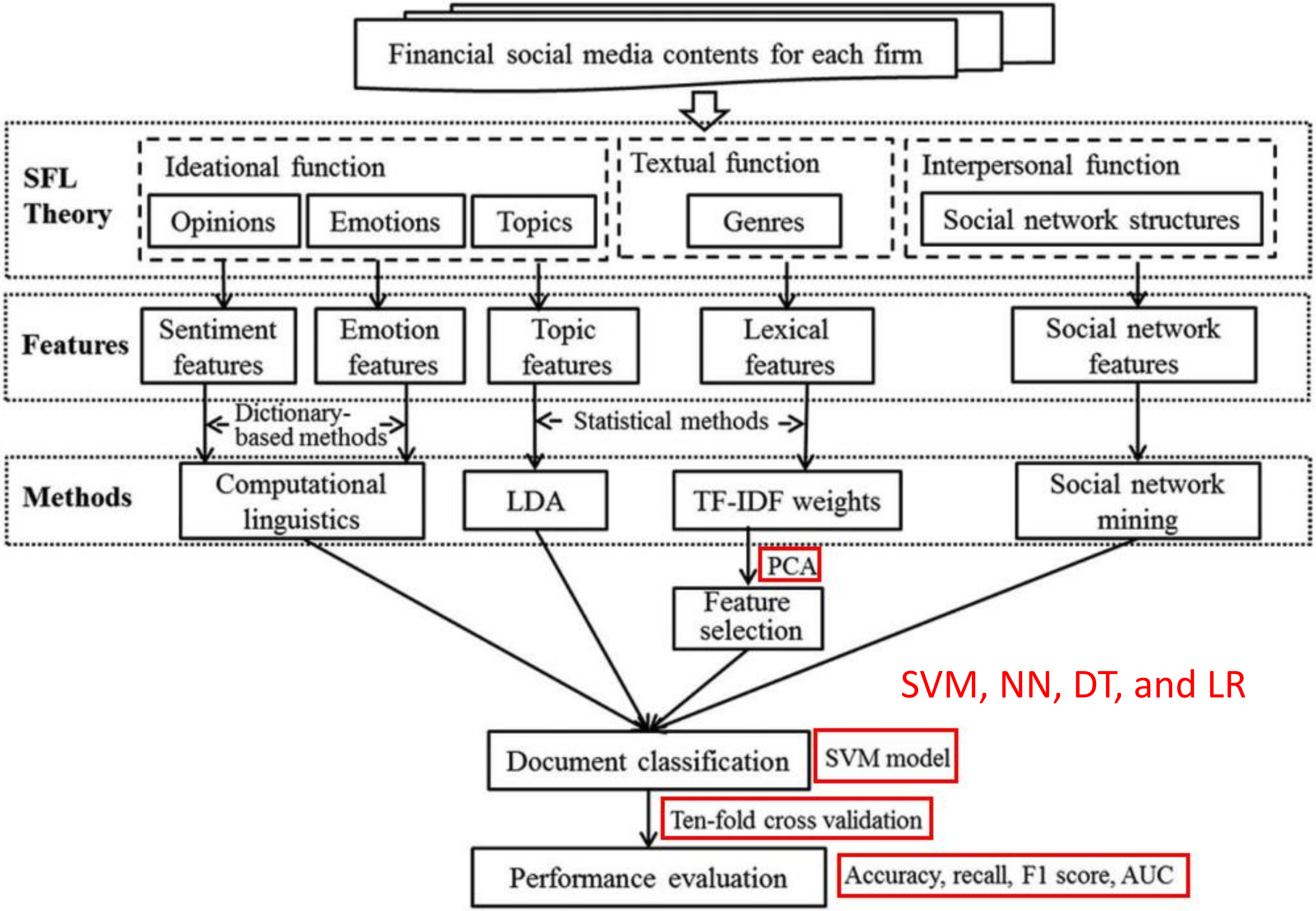
It is generally represented by social structure that can be built through the reply-to relationships between messages.

Social structural characteristics can be used to delineate deception in computer mediated communication.

Method:

Type	Feature and measurement	No. of features
Social interaction structure	Number of Analysis reports (AR), Breaking news (BN), or StockTalk messages (SM)	3
	Number of comments to AR, BN, or SM	3
	Numbers of distinctive authors for AR, or SM	2
	Numbers of AR or SM per author	2
	Number of followers	1

An SFL-Based Framework



Data

- Financial social media data source: SeekingAlpha
  - insights are provided by investors and industry experts from the buy-side rather than the sell-side
  - For each firm, there are five types of topic discussions: **Analysis reports, Breaking news**, Earning call transcripts, **StockTalk**, and Videos.

- Financial ratios and the textual contents of the MD&A section from the annual financial statements are from the Compustat and the EDGAR database.
  - Financial ratios: 12 annual financial ratios, 24 industry-collaboration contextual features, 24 industry-competition contextual features, and 24 organization contextual features

Sample Selection

Distinct companies	Number
Companies with accounting misconducts in Dechow et al. [19] data set	936
Less: companies with only quarterly fraudulent events	132
Subtotal (companies with annual fraudulent events)	**804**
Less: companies with auditor, bribes, disclosure, no dates, and other issues	38
Subtotal (companies with annual corporate fraud)	**766**
Less: Companies that cannot be found in SEC EDGAR database	111
Less: Companies that cannot be found in Compustat database	38
Less: Companies that cannot be found in SeekingAlpha	343
Less: Financial companies: Banks & Insurance (SIC 6000-6999)	103
Less: Companies' financial data in fraud years cannot be found in Compustat database	22
Less: Companies that are disclosed before the establishment of SeekingAlpha	29
Less: Companies that do not have enough social media data	56
Total	**64**

Sample Selection

- Oversampling strategy (1:1 ratio for fraudulent and nonfraudulent firms):

1. find a direct match on the basis of the fraud year, size, and industry
2. each firm should also have enough textual data on SeekingAlpha
3. randomly chose if many firms meet the selection criteria

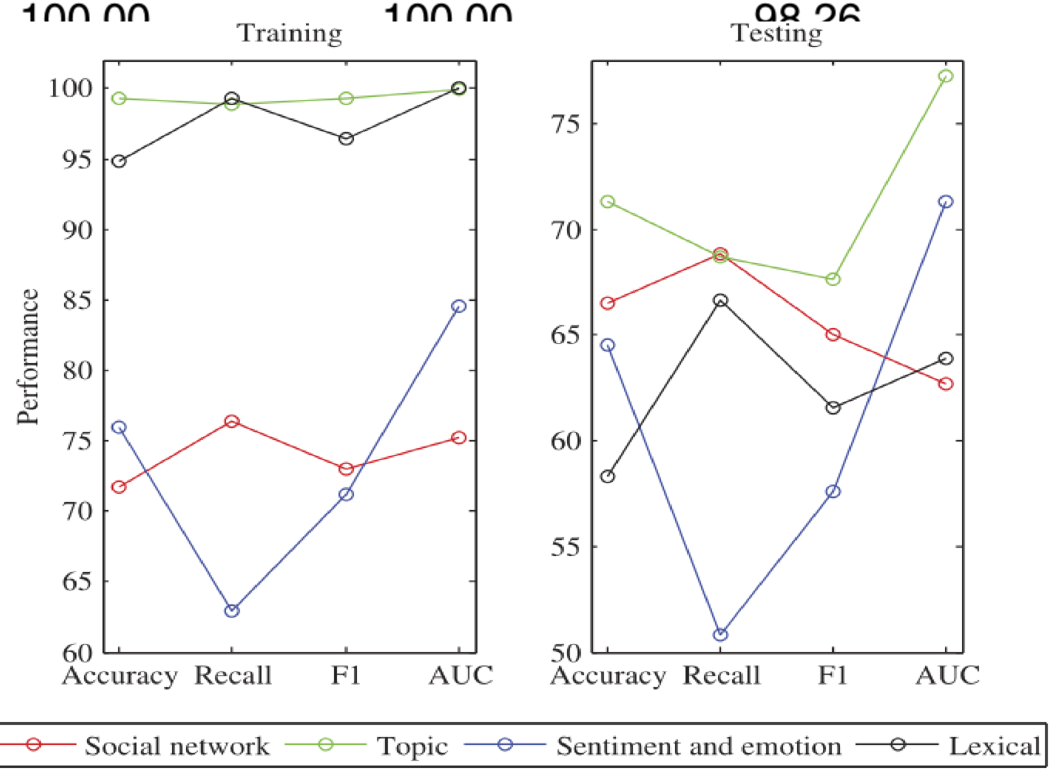
➢ Final data set includes 64 fraudulent firms together with a corresponding 64 matched nonfraudulent firms.

Dataset (128 firms)	No. of Analysis reports	No. of Breaking news	No. of StockTalk messages	No. of sentences	No. of words	No. of financial ratios
Social media data	3,981 (31.10)	2,251 (17.59)	1,672 (13.06)	184,356 (1,440.28)	2,613,362 (20,416.89)	—
MD&A data	—	—	—	92,712 (724.31)	902,940 (7,054.22)	—
Financial ratios	—	—	—	—	—	84

Is the financial social media data useful?

- Fraud Detection Using Only Social Media Data: 2 sentiment features, 8 emotion features, and 11 social network features

		Average accuracy	Average recall	Average F1 score	Average AUC
SVM	Training	99.66	99.50	99.66	99.94
	Testing	75.50	81.56	76.50	86.32
NN	Training	100.00	100.00	100.00	98.26
	Testing	63.17			
DT	Training	98.52			
	Testing	63.10			
LR	Training	50.27			
	Testing	54.50			



Is the financial social media data useful?

		Average accuracy	Average recall	Average F1 score	Average AUC
SVM	Training	99.66	99.50	99.66	99.94
	Testing	75.50	81.56	76.50	86.32
NN	Training	100.00	100.00	100.00	98.26
	Testing	63.17	68.05	62.18	53.71
DT	Training	98.52	98.30	98.52	96.44
	Testing	63.10	66.54	64.93	43.34
LR	Training	50.27	87.04	59.96	46.42
	Testing	54.50	87.75	60.98	43.70

Using Only Social Media Data

		Average accuracy	Average recall	Average F1 score	Average AUC
SVM	Training	99.39	98.77	99.37	99.96
	Testing	56.17	77.74	63.37	49.29
NN	Training	76.58	69.60	74.72	73.09
	Testing	48.83	42.39	43.71	41.75
DT	Training	97.41	96.75	97.37	95.34
	Testing	41.17	42.08	40.09	36.02
LR	Training	54.69	60.31	56.88	51.94
	Testing	54.67	60.00	54.54	43.58

Using Only Financial Ratios

		Average accuracy	Average recall	Average F1 score	Average AUC
SVM	Training	97.29	97.01	97.28	99.50
	Testing	66.67	66.02	64.25	69.82
NN	Training	100.00	100.00	100.00	98.26
	Testing	66.33	62.19	64.69	52.09
DT	Training	99.14	98.81	99.10	97.57
	Testing	52.78	50.83	54.98	54.14
LR	Training	100.00	100.00	100.00	98.26
	Testing	70.33	71.90	70.10	58.96

Using Only Language-based Features from MD&A Contents

Is there any incremental value?

Table 10. Performance of SVM Classifier Using Combined Features

	SVM		Average accuracy	Average recall	Average F1 score	Average AUC
SVM	Financial ratios	Training	99.39	98.77	99.37	99.96
		Testing	56.17	77.74	63.37	49.29
	Financial ratios and language-based features	Training	98.71	98.60	98.72	99.83
		Testing	70.83	68.54	69.31	71.78
	Fully combination of features	Training	100.00	100.00	100.00	100.00
		Testing	80.00	83.04	79.80	85.03
NN	Financial ratios	Training	76.58	69.60	74.72	73.09
		Testing	48.83	42.39	43.71	41.75
	Financial ratios and language-based features	Training	100.00	100.00	100.00	98.26
		Testing	62.33	69.48	63.91	50.75
	Fully combination of features	Training	100.00	100.00	100.00	98.26
		Testing	66.17	79.96	69.80	55.07
DT	Financial ratios	Training	97.41	96.75	97.37	95.34
		Testing	41.17	42.08	40.09	36.02
	Financial ratios and language-based features	Training	97.99	98.60	98.03	96.33
		Testing	52.78	52.92	46.88	47.25
	Full combination of features	Training	98.28	99.29	98.36	96.67
		Testing	52.38	49.18	45.58	39.35
LR	Financial ratios	Training	54.69	60.31	56.88	51.94
		Testing	54.67	60.00	54.54	43.58
	Financial ratios and language-based features	Training	54.33	61.34	56.93	51.22
		Testing	53.00	56.35	53.09	48.97
	Full combination of features	Training	98.28	99.29	98.36	96.67
		Testing	52.38	49.18	45.58	39.35

Will be influenced by rumors in social media?

- Too many false rumors in SeekingAlpha can cause issues in data quality and lead to imprecise classification of fraudulent and nonfraudulent firms.

		Average accuracy	Average recall	Average F1 score	Average AUC
SVM	Training	99.66	99.50	99.66	99.94
	Testing	75.50	81.56	76.50	86.32

SVM		Average accuracy	Average recall	Average F1 score	Average AUC
No rumor only	Training	98.97	98.96	98.97	99.78
	Testing	78.33	74.20	76.08	84.13
No leaked information only	Training	95.69	93.45	95.49	99.25
	Testing	76.17	73.27	74.11	84.59
No rumor and no leaked information	Training	98.19	97.76	98.15	99.61
	Testing	77.00	72.26	74.94	80.94

Other Generalizability Check

- We extend Dechow et al.'s study period to December 31, 2014, carefully examine the 127 new AAERs and finally found 13 of them fraudulent (7). The average accuracy, recall, F1 measure, and AUC of our model on this holdout sample are 71.43 percent, 57.14 percent, 66.67 percent, and 69.39 percent, respectively.
- We assemble a new data set from another social media platform, Yahoo Finance.

Conclusion

- This study used social media data from financial platforms and proposed a text analytic framework, rooted in the SFL theory, which aims to extract signals/cues from social media data to detect early signs of fraud.
- We not only demonstrate the efficacy of social media features for fraud detection but also verify that a social media-based method can supplement existing corporate fraud detection approaches.

想法

- 没有提及提取社交媒体数据的时间窗口
- 谣言对模型影响的检验是一种事后的检验，无法在事前进行评断
- SFL-Based Framework预测股票收益