MICHIGAN

ROSS SCHOOL OF BUSINESS

Financial Analysis and Forecasting Using Textual Data

Feng Li
University of Michigan

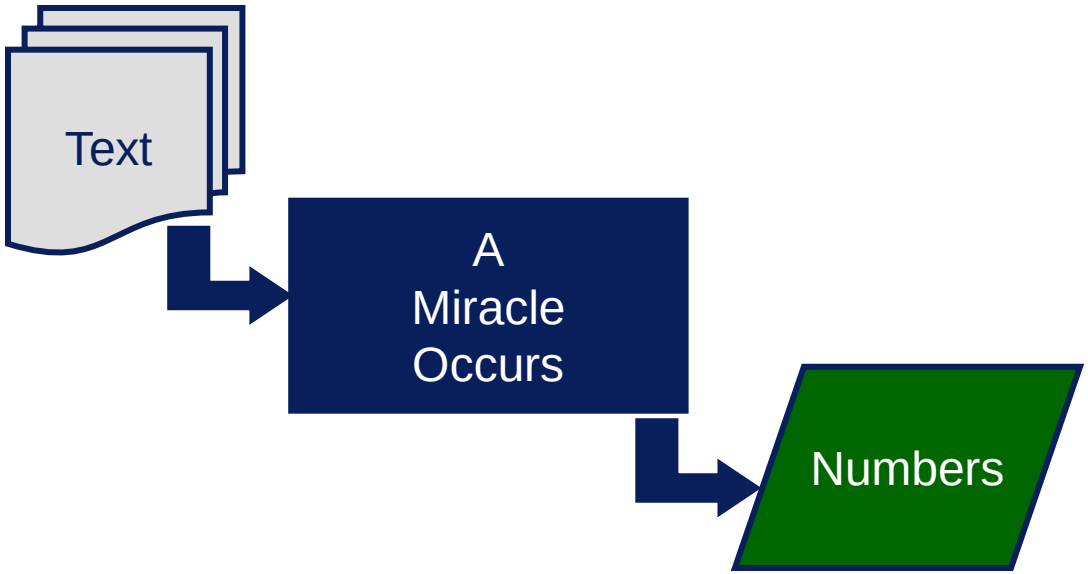
CARE conference
Miami, Florida
April 10th, 2010

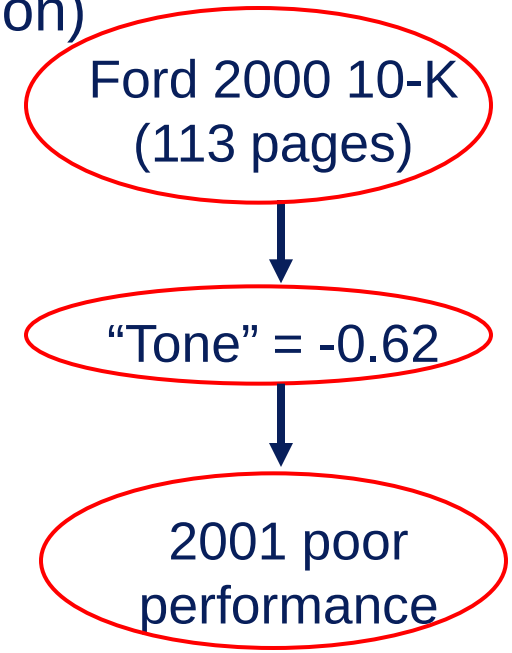
Roadmap

- Why textual information
- Different approaches
- Future research opportunities

Why text? (1)

- Important information source
  - ~ Typical financial statements: 200+ numbers, 30,000+ words
  - ~ Transactions ➔ net income (aggregation)
  - ~ Text ➔ numeric variables (data reduction)





Why text? (2)

- Data generating function
  - ~ Notes to the financial statements
  - ~ E.g., sales revenue increases, but revenue recognition method changed
- Can be forward-looking compared with many other fundamental signals

> "We have incurred significant losses since our inception, and we expect to continue to incur net losses for the foreseeable future."

Understanding text

- Not really something new
  - ~ E.g., Bible gospels authorship study
  - ~ E.g., using footnote to understand financial numbers
  - ~ The question is: Can we do it quantitatively using large sample?

- Inter-disciplinary: Linguists, statisticians, computer scientists, psychologists; content analysis, text mining, computational linguistics, natural language processing

- There is relatively little empirical research on financial forecasting using textual data

Sources of textual data

- Potentially interesting for financial forecasting
  - ~ SEC filings
  - ~ Earnings releases
  - ~ Conference calls
  - ~ Financial news
  - ~ Analyst reports

Two ways of using textual data

- Textual information as main variable
  - ~ But … 1 number might >= 1000 words

- Textual information as contextual variable
  - ~ Combine textual information with other variables of interest (earnings, prices…)
  - ~ E.g., accrual anomaly as a function of …

Understand text: parsimonious approach

- Frequency count of specific words
- Keyword extraction
  - ~ Extract key sentences and strip away anything unimportant
  - ~ Look for particularly important keywords, nouns and noun phrases that recur frequently in a text

Example: GE's "greatometer" (*Scott Davis, Morgan Stanley's lead GE analyst)*

- In 2002 Q3 call, Messrs. Immelt and Sherin said "great" more than 20 times.
- 2005 Q2, 70 times (GE shares rose 37%)
- 2006 Q3, 37 times (GE stock fell 10%)
- 2007 Q4, 80 incidences of greatness, including the "great company," the "great quarter," the "great momentum" and the "great risk management." (GE shares rose to $40.16.)

Example: Manager heuristics (self-attribution bias)

American International Group, Inc. 2006 Annual Report

- "Solid execution of our strategies and the absence of significant catastrophes contributed to our outstanding results in 2006. Around the world and across all of our business segments we are capitalizing on growth opportunities, using our business diversity and matrix management structure to respond quickly to customer needs."

Example: Manager heuristics (self-attribution bias)

American International Group, Inc. 2008 Annual Report

- "AIG reported that the continued severe credit market deterioration, particularly in mortgage-backed securities, and charges related to ongoing restructuring activities, contributed to a record net loss for the fourth quarter of $61.7 billion, or $22.95 per diluted share, compared to a 2007 fourth quarter net loss of $5.3 billion, or $2.08 per diluted share."

Empirical evidence

- Evidence shows that this approach might work reasonably well.
- Li (2010): managers' self-attribution bias is associated with operation turnover rate, return volatility etc.
- Chatterjee and Hambrick (2007): CEO narcissism (measured using simple word count) has implications for company strategy and performance.

Understand text: dictionary approach

- Mapping algorithm
  - ~ For e.g., dictionaries written by psychologists
    - DICTION
    - General Inquirer
    - LIWC

Example

- “It is a cloudy day and the stock market is not doing well.”

General Inquirer output

tag	N	%	words
Pos	1	7.69	WELL#2=1
Pstv	1	7.69	WELL#2=1
Virtue	1	7.69	WELL#2=1
Strng	1	7.69	DO#1=1
Actv	1	7.69	DO#1=1
Negate	1	7.69	NOT=1
Econ*	1	7.69	MARKET#1=1
Means	1	7.69	STOCK=1
Work	1	7.69	DO#1=1
Time*	1	7.69	DAY=1
PLACE	1	7.69	MARKET#1=1
Social	1	7.69	MARKET#1=1
PRON	1	7.69	IT=1
DAV	1	7.69	DO#1=1
SV	2	15.38	ARE#1=1 ARE#2=1

Pros and cons of dictionary approach

- Pros
  - ~ Easy
- Cons
  - ~ Tailored dictionary was not readily available
  - ~ “Downgrade” is not a negative word in GI
  - ~ Prior ignored
  - ~ Context ignored
- Essentially it’s a word-level analysis

Performance of the dictionary-based measures

Information Content of Tone Measures Based on the Dictionary Approach

COEFFICIENT	(1) EARN(t+1)	(2) EARN(t+1)	(3) EARN(t+1)	(4) EARN(t+1)	(5) EARN(t+1)	(6) EARN(t+1)
DICTION_POS	-0.000					
	(-1.13)					
DICTION_NEG	-0.000**					
	(-2.03)					
DICTION_TONE		0.000				
		(1.06)				
GI_POSITIV			-0.002***			
			(-9.05)			
GI_NEGATIV			-0.001**			
			(-2.57)			
GI_TONE				-0.001***		
				(-5.72)		
LIWC_POSEMO					-0.002***	
					(-5.94)	
LIWC_NEGEMO					-0.001	
					(-1.15)	
LIWC_TONE						-0.001***
						(-4.49)
Observations	109749	109749	113802	113802	113207	113207
R^2	0.36	0.36	0.36	0.36	0.36	0.36

Negative or insignificant correlations with future earnings

Low correlations with current earnings

Correlations Between Dictionary Tone Measures and Current Earnings and Fog

	TONE	DICTION_TONE	GI_TONE	LIWC_TONE
EARN	0.15	0.03	-0.03	-0.02
	(0.00)	(0.00)	(0.00)	(0.00)

Dictionary tailored for financial setting

- Henry and Leone (2010)

Panel A. Regression of Abnormal Returns on Unexpected Earnings and Tone Scores, controlling for Loss-making Companies

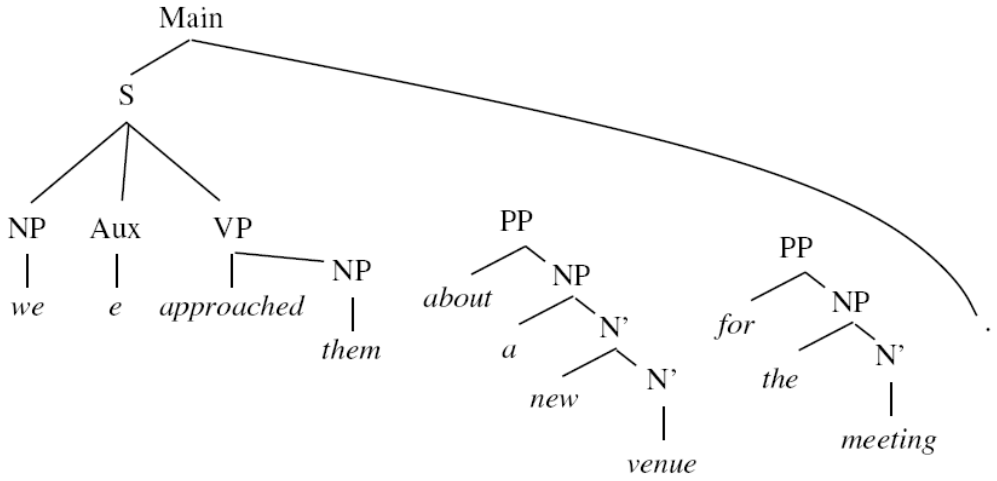
	Coefficient (t-Stat)	Coefficient (t-Stat)	Coefficient (t-Stat)
INTERCEPT	-0.011*** (-4.330)	0.000 (-0.083)	-0.006* (-2.124)
UE	0.208*** (10.941)	0.219*** (11.390)	0.217*** (11.320)
SIZE	0.001*** (3.042)	0.001*** (2.723)	0.001*** (3.646)
LOSS	-0.012*** (-8.416)	-0.013*** (-8.453)	-0.013*** (-9.312)
FD_SCORE	0.024*** (12.289)		
DICTION_SCORE		0.007*** (4.637)	
GI_SCORE			0.014*** (4.633)
Year Fixed Effects	Yes	Yes	Yes
Adjusted R2	3.10%	2.40%	2.40%

Understand text: statistical approach

- Train on some data → learn → predict
  - Naïve Bayesian (Fast; decent results)
  - SVM (Support vector machine)
- Let the data decide the pattern
- Danger: over-fitting the data
  - ~ Limited success of neural networks in financial forecasting
- Important: Cross-validation tests to guard against over-fitting

Understand text: more structured approach

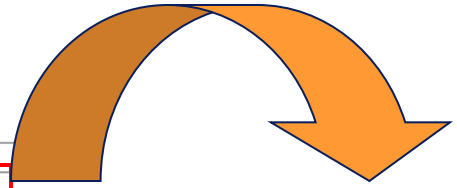
- Par-of-speech (POS) tagging
  - ~ Rule-Based tagging (Voutilainen 1995)
  - ~ Stochastic (e.g., Hidden Markov Model) tagging (Brants 2000)
  - ~ Transformation-based tagging (Brill 1995)



Promising direction

Consolidated Statements of Income

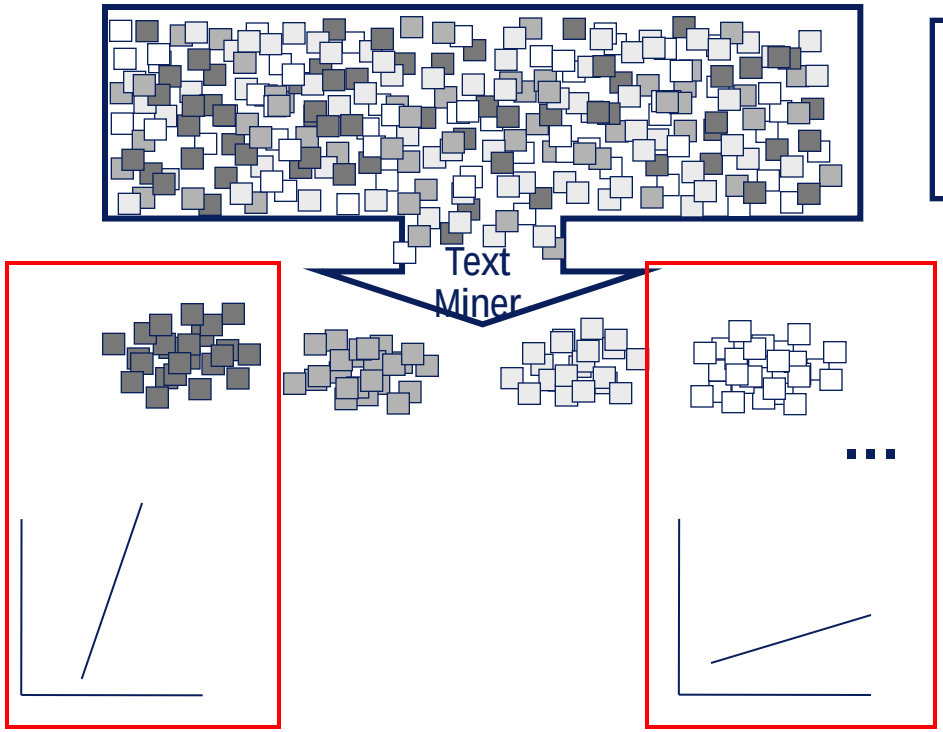
(In millions, except per share amounts)	2009
REVENUES	$35,497
OPERATING EXPENSES:	
Salaries and employee benefits	13,767
Purchased transportation	4,534
Rentals and landing fees	2,429
Depreciation and amortization	1,975
Fuel	3,811
Maintenance and repairs	1,898
Impairment and other charges	1,204
Other	5,132
	34,750
OPERATING INCOME	747

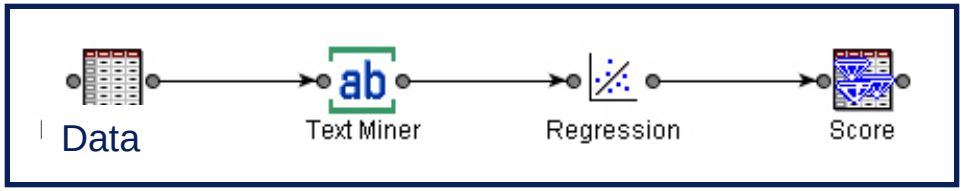


Search for all the sentences that have “sales” or “revenues” as NP, and then extract the VP, ADJP, and ADVP in these sentences for further analysis (e.g., “increase”)

Textual information as a tool to separate heterogeneous data

- Document classification and clustering





Parse

Transform

Cluster

Example

- Document similarity using K-means
  - ~ Over-time comparison
  - ~ Cross-sectional comparison
    - – E.g., Revenue recognition policy compared with the same firm last year, with firms in the same industry etc.

Summary

- This is an exciting area. A lot more can be done.
- More structured and rigorous approach will be more desirable and fruitful for future research.