Predicting Financial Market Activities Using Aggregated Industry-Level Text Data

Hsin-Min Lu
盧信銘
Department of Information Management
National Taiwan University

1

Who am I

- Hsinmin Lu [盧信銘]
  - Ph.D. in Management Information Systems, University of Arizona
  - Master of Arts in Economics, National Taiwan University
  - Bachelor of Business Administration, National Taiwan University
- Research Interests
  - Text and Data Mining
  - Business Intelligence
  - Empirical Finance
  - Applied Econometrics
  - Medical Informatics

2

Agenda

- Introduction
- Literature Review
- Research Gaps and Questions
- Design Framework
- Research Testbed
- Experimental Results
- Conclusions

3

INTRODUCTION

4

Introduction

- Predictive modeling
  - Valuable for organizations in rapidly changing environment
  - An important feature for enterprise information systems
  - Essential for information-intensive industries, e.g., financial industry

5

Introduction (Cont'd.)

- Financial market activities can be predicted using structured data
  - E.g., stock returns (Fama French 1992; Haugen and Baker 1996)
  - Relevant "soft information" was discarded
    - Text data from news articles, financial reports, and investor forum postings
- Recent research responded to this limitation
  - Stock price 20-min after news release is predictable (Schumaker and Chen 2009)
  - Large volatility movement after news release is also predictable (Groth and Muntermann, 2010)
  - Trading strategies based on news coverage can generate abnormal returns (Fang and Peress, 2009)

6

Introduction (Cont'd.)

- Most studies investigated the impact of text data on individual firms
  - Investor do not relate message to other same-industry firms
- Our study investigate the value of aggregated industry-level news data
  - Triangulating text data with financial market activity measures
  - Conduct simulated trading and predictive regression
  - Focusing on the impact on returns, volatility and trading volume

7

LITERATURE REVIEW

8

Literature Review

- Text representations for financial market activity prediction
- Predicting financial market activities using text data

Text Representations for Financial Market Activity Prediction

- General-purpose text representation
  - Bag-of-word (BOW) (Fung et al. 2002; Groth and Muntermann 2010)
    - Converts a document into a long numeric vector
    - Each dimension represent a term in the document
    - TF-IDF is often used
  - Subset of terms, e.g., named entity, noun phrases, proper nouns (Schumaker and Chen, 2009)
  - Convenient to use
  - Not designed to assess the statistical significance of individual dimension

Text Representations for Financial Market Activity Prediction (Cont'd.)

- Selected aspects of text data
  - Message volume
  - Sentiment
  - Degree of risk-relevance

Text Representations for Financial Market Activity Prediction (Cont'd.)

- Message volume: # of posting/news articles (Das and Chen, 2007)
  - A coarse measure
  - Proxy information flow (Berry and Howe 1994)
  - Appropriate if
    - Content has been pre-processed, or
    - Content is not important

Text Representations for Financial Market Activity Prediction (Cont'd.)

- Sentiment: positive and negative opinions, emotions, and evaluations embedded in text data (Wiebe et al., 2005)
  - Dictionary based sentiment identification
    - Identify polarity words in news articles (Tetlock 2007) and financial reports (Kothari et al. 2009)
    - Suitable for well-written documents
  - Supervised learning approaches (Antweiler and Frank 2004; Das and Chen 2007)
    - Naïve Bayes and SVM
    - Lower the negative effect of slangs, abbreviations, typos, and grammatical errors in forum postings

13

Text Representations for Financial Market Activity Prediction (Cont'd.)

- Sentiment analysis results need to be aggregated and transformed for predictive modeling
- Various measures have been proposed:

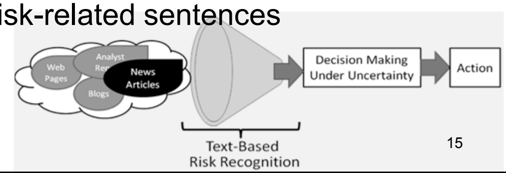
$$Senti1_t = \frac{M_t^{POS} - M_t^{NEG}}{M_t^{POS} + M_t^{NEG}} \qquad Senti2_t = ln\frac{1 + M_t^{POS}}{1 + M_t^{NEG}}$$

$$DISAG = |1 - |\frac{M_t^{POS} - M_t^{NEG}}{M_t^{POS} + M_t^{NEG}}|| \qquad AG = 1 - \sqrt{1 - Senti1_t^2}$$

14

Text Representations for Financial Market Activity Prediction (Cont'd.)

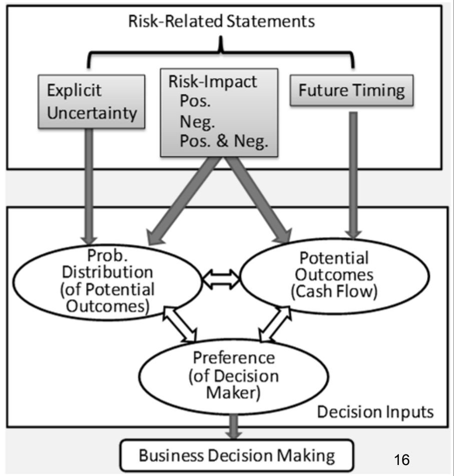
- Risks
  - The potential events and trends that may impact a business's growth trajectory and shareholder value (COSO 2004; Slywotzky and Drzik 2005)
  - Often conveyed in qualitative text descriptive
  - Tracking and monitoring can be costly
  - Lu et al. (2009) proposed a design framework to recognize risk-related sentences



15

Text Representations for Financial Market Activity Prediction (Cont'd.)

- Risk-related statements contribute to the inputs for decision making under uncertainty
- The contributions are captured by:
  - Future Timing
  - Explicit Uncertainty
  - Risk Impact



16

Text Representations for Financial Market Activity Prediction (Cont'd.)

- Future Timing (FT): Whether the primary content of a sentence is about future events or states
- Explicit Uncertainty (EU): Whether this sentence contains explicit accounts of doubt or unreliability
- Risk Impact (RALL): Whether a sentence contains information affecting decision makers' beliefs over a firm's future cash flow
  - RP & RN: the direction of impact

No	Sentence	RALL	RP	RN	FT	UC
1	While many analysts had predicted the market for ICDs would grow about 20% a year due to an aging population, many now forecast only single- digit percentage growth for the year (Wall Street Journal, Oct. 19, 2006).	√		√	√	√
2	Many personal-computer applications send hundreds, even thousands of messages back and forth before completing a task such as transferring a file (Wall Street Journal, Apr. 27, 2004).					

17

Text Representations for Financial Market Activity Prediction (Cont'd.)

- Documents containing more risk-related sentences should be more informative to investors
  - Can be measured by the proportion of risk-related sentences → Degree of risk-relevance
  - Few studies have investigated it's economic impact

18

Predicting Financial Market Activities Using Text Data

- Three financial market activity measures were considered in previous studies:
  - Returns
  - Volatility
  - Trading volume

19

Predicting Financial Market Activities Using Text Data (Cont'd.)

- Returns
  - Predictive models including both text and numeric data outperformed models considering only numeric data (Schumaker and Chen, 2009)
  - Increase in investor forum postings predicted lower future returns (Antweiler and Frank, 2004)
  - Trading strategies based on news coverage can generate statistically significant abnormal returns (Fang and Peress, 2009)
  - News sentiment can predict short-term market return (Tetlock 2007)

20

Predicting Financial Market Activities Using Text Data (Cont'd.)

- Volatility
  - A good proxy of risks
  - The relationship between future volatility and the release of mandatory corporate disclosure can be learned (Groth and Muntermann, 2010)
  - Sentiment in financial reports and news articles is associated with volatility (Kothari et al. 2009)
  - Investor forum sentiment can predict return volatility (Antweiler and Frank 2004)

21

Predicting Financial Market Activities Using Text Data (Cont'd.)

- Trading volume
  - Unusually high or low news pessimism predicts higher daily market trading volume (Tetlock, 2007)
  - Higher level of message posting in investor forums predicts higher trading volume (Antweiler and Frank, 2004)
  - Disagreement in forum messages also predicts higher trading volume (Antweiler and Frank, 2004)

22

RESEARCH GAPS AND QUESTIONS

23

Research Gaps

- Most studies focus on the interaction between text data and financial markets at firm- and market-level
  - Few studies have investigated the predictive value of text data at industry-level
- Risk-related information is presumably important for investors
  - Few studies have investigate its impact on financial markets

24

Research Questions

- Can we predict financial market activities using aggregated industry-level text data?
  - Generate abnormal returns
  - Predict future volatility
  - Predict future trading volume
- Is risk-related information valuable for financial market activities prediction?

25

DESIGN FRAMEWORK FOR

SUPPORTING PREDICTIVE MODELING FOR FINANCIAL MARKET ACTIVITIES

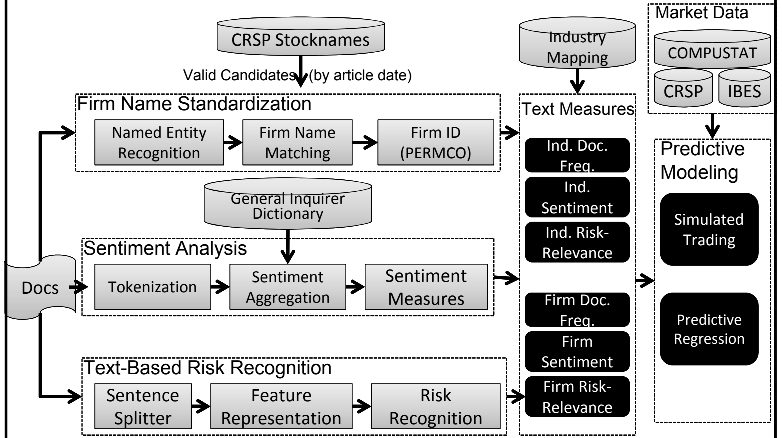
26

Design Rationale

- Focusing on documents that can be linked to public firms explicitly
- Consider selected aspects of text data
  - Message volume, sentiment, and degree of risk-relevance
  - May be extended to include general-purpose text representation later.
- Aggregate text measures at firm- and industry-level
- Evaluation:
  - Simulated trading and predictive regression

27

Design Framework



28

Firm Name Standardization

- Extract institution names using a named entity recognition component
- Match against known publicly traded firms
  – The “CRSP stocknames” table
- Tight-to-lose approach
  – Full string matching
  – Truncated string matching
  – Handles variation caused by punctuation marks and acronyms
  – Outputs PERMCO, a standard numeric ID for public firms

29

Sentiment Analysis

- Word-level analysis
  – Tokens preprocessed using a WordNet-based morphological analyzer
- Identify positive and negative words using General Inquirer Dictionary (Tetlock, 2007; Fang and Peress, 2009)
  – Cf. http://www.wjh.harvard.edu/~inquirer/
  – Records # of positive and negative words

30

Text-Based Risk Recognition

- Identify risk-related sentences
- The risk recognition model was trained using 2539 manually tagged sentences
  – Extracted from the WSJ
  – Elastic-net logistic regression
  – Accuracy=69.4%; F-measure = 68.9%
- Recorded # of risk-related sentences and total # of sentences

31

Firm-Level Text Measures

- Document frequency: # of docs that can be linked to a firm (during a time period)
- Sentiment: $Senti1 = \frac{M^{POS} - M^{NEG}}{M^{POS} + M^{NEG}}$
- Degree of risk-relevance

$$RALL = \frac{\#\ of\ RiskRelated\ Sentences}{\#\ of\ Sentences}$$

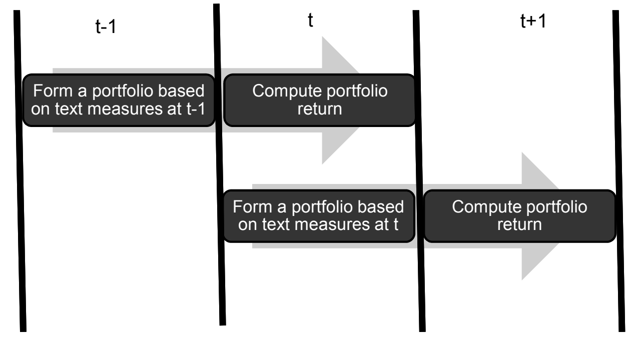
32

Industry-Level Text Measures

- 49 industries according to Fama and French (1997)
  - E.g. Telcm Communication includes
    - 4800-4800 Communications
    - 4810-4813 Telephone communications
    - 4820-4822 Telegraph and other message communication
    - 4830-4839 Radio-TV Broadcasters
    - … (truncated)
- Value-weighted firm-level measures
  - Weighting reflects the importance of firms

33

Simulated Trading



34

Simulated Trading (Cont'd.)

- Text measures considered
  - Industry-level message volume
  - Industry-level sentiment
  - Industry-level degree of risk-relevance
- Portfolio formation strategy
  - Following Fang and Peress (2009)
  - Sort industries by a text measure
  - Split industries into 3 groups by 30% - 70% percentile
  - Form equally weighted zero-investment portfolio buy buying top-30 percentile industries and selling bottom-30 percentile industries

35

Simulated Trading (Cont'd.)

- Trading strategy evaluation
  - Raw trading return (t test)
  - Abnormal return (adjusting for Fama-French risk factors) (Fang and Peress, 2009)

$$tr_t = \alpha + \beta_1 Mktrf_t + \beta_2 SMB_t + \beta_3 HML_t + e_t$$

- tr: trading return
- Mktrf: excess market return
- SMB: Small-cap stock return minus large-cap stock return
- HML: high book-to-market ratio stock returns minus low book-to-market ratio stock returns

36

Predictive Regression: Volatility

- Regress future firm volatility on industry-level text measures
- Controlled for
  - Firm-level text measures
  - Lag volatility, lag return, lag trading volume, firm size, book-to-market ratio, proportion of individual ownership, analysts' coverage, and analysts' earnings prediction dispersion (Andersen 1996, Fama and French 1993, Fang and Peress 2009, Kothari 2001)
- Significant coefficients indicate the effect of industry-level text measures

37

Predictive Regression: Trading volume

- Regress future firm trading volume on industry-level text measures
- Controlled for
  - Firm-level text measures
  - Lag volatility, lag return, lag trading volume, firm size, book-to-market ratio, proportion of individual ownership, analysts' coverage, and analysts' earnings prediction dispersion (Andersen 1996, Fama and French 1993, Fang and Peress 2009, Kothari 2001)
- Significant coefficients indicate the effect of industry-level text measures

38

Research Testbed

- Text data: the Wall Street Journal (WSJ)
  - Best selling newspaper in the U.S. (as of March, 2010)
  - January 1985 and May 2008
  - 1,134,332 news articles
- Financial market data
  - CRSP (daily and monthly stock prices)
  - Compustat (financial reports)
  - IBES (Analysts' forecast)
- 1,274,711 firm-months

39

EXPERIMENTAL RESULTS

40

Simulated Trading Results

	Trading Return average return (t-value)	Abnormal Return average return (t-value)
Message Volume	0.0012 (0.92)	0.0019 (1.47)
Sentiment	0.0012 (0.88)	0.0019 (1.42)
Degree of Risk-Relevance	**0.0025**** (1.99)	**0.0034***** (2.64)

***, **, * indicate statistical significance at the 0.001, 0.05, and 0.1 levels, respectively.

Simulated Trading Results

- Simulated trading based on industry-level degree of risk-relevance generated significant abnormal return
  - 0.0034 per month ➔ 4.08% per year
- Industry-level message volume and sentiment did not generate significant abnormal returns

Predictive Regression: Future Volatility

	Baseline		Full Model	
	Estimate	t-value	Estimate	t-value
Intercept	**-1.400*****	-24.754	**-1.430*****	-23.094
Firm - Msg Volume			**0.008*****	6.065
Firm - Senti1			**-0.127*****	-14.243
Firm – Deg. Risk			**-0.041*****	-3.088
Ind. – Msg Volume			**0.001*****	3.266
Ind. – Senti1			-0.102	-1.031
Ind. – Deg. of Risk			**0.417*****	5.863
Log Volatility	**0.620*****	80.225	**0.617*****	81.176
Log Size	**-0.192*****	-30.881	**-0.196*****	-30.116
Log Volume	**0.098*****	24.189	**0.095*****	24.620
Log BM	**-0.067*****	-17.247	**-0.066*****	-16.857
Ret	**-0.586*****	-12.502	**-0.580*****	-12.642
Indv. Own	**0.082****	2.997	**0.075****	2.971
Log AnalyCover	0.003	0.308	0.011	1.282
Log AnalySD	**0.063*****	5.433	**0.062*****	5.551
ADJ-RSQ	0.613		0.615	

***, **, * indicate statistical significance at the 0.01, 0.05, and 0.1 levels, respectively.

Predictive Regression: Volatility (Cont'd.)

- Industry-level text data predict future volatility
  - Higher message volume → Higher future volatility
  - Higher degree of risk-relevance → Higher future volatility
  - Sentiment has no effect on future volatility
- Firm-level text data also impact future volatility
  - Higher message volume → Higher future volatility
  - Higher sentiment → Lower future volatility
  - Higher degree of risk-relevance → Lower future volatility
- Degree of risk-relevance: the direction of impact is different on firm- and industry-level.
  - May be caused by the competition among same-industry firms

Predictive Regression: Future Trading Volume

	Baseline		Full Model	
	Estimate	t-value	Estimate	t-value
Intercept	**0.522*****	10.04	**0.490*****	8.65
Firm - Msg Volume			**0.002****	2.03
Firm - Senti1			**-0.077*****	-13.97
Firm – Deg. Risk			**-0.033*****	-3.85
Ind. – Msg Volume			**0.001*****	2.64
Ind. – Sentiment			-0.100	-1.25
Ind. – Deg. of Risk			**0.236*****	4.33
Log Volume	**0.876*****	222.89	**0.875*****	226.76
Log Volatility	-0.006	-1.10	-0.007	-1.25
Log Size	**0.048*****	11.39	**0.048*****	10.94
Log BM	**-0.055*****	-19.54	**-0.053*****	-18.65
Ret	**-0.156*****	-5.14	**-0.153*****	-5.14
Indv. Own	**-0.201*****	-12.04	**-0.197*****	-12.79
Log AnalyCover	**0.054*****	8.29	**0.06*****	9.96
Log AnalySD	0.017	1.49	0.017	1.50
ADJ-RSQ	0.90		0.90	

***, **, * indicate statistical significance at the 0.01, 0.05, and 0.1 levels, respectively.

Predictive Regression: Trading Volume (Cont'd.)

- Industry-level text data predict future trading volume
  - Higher message volume → Higher future trading volume
  - Higher degree of risk-relevance → Higher future trading volume
  - Sentiment has no effect on future volatility
- Firm-level text data also impact future trading volume
  - Higher message volume → Higher future trading volume
  - Higher sentiment → Lower future volatility
  - Higher degree of risk-relevance → Lower future volatility
- Degree of risk-relevance: the direction of impact is different on firm- and industry-level.

CONCLUSIONS & CONTRIBUTIONS

Conclusions

- We developed a design framework to study the effect of aggregated industry-level text data
- Our experimental results show that
  - Trading strategy based on industry-level risk-relevance can generate significant abnormal returns
  - Industry-level message volume can predict future volatility and trading volume
  - Industry-level sentiment has no effect on future volatility and trading volume

Contributions

- To the best of our knowledge, this is the first study that
  - Investigate the economic effect of aggregated industry-level text data
  - Investigate the effect of risk-related information in news articles
- Future works
  - Refinement of various text measures
  - Study the effect at higher frequency
  - Text data from different sources

49

Questions

50