大连理工大学
信息检索研究室
Information Retrieval Laboratory of DUT

News impact on stock price return via sentiment analysis

赵明珍

2015年4月23日

Paper Introduction

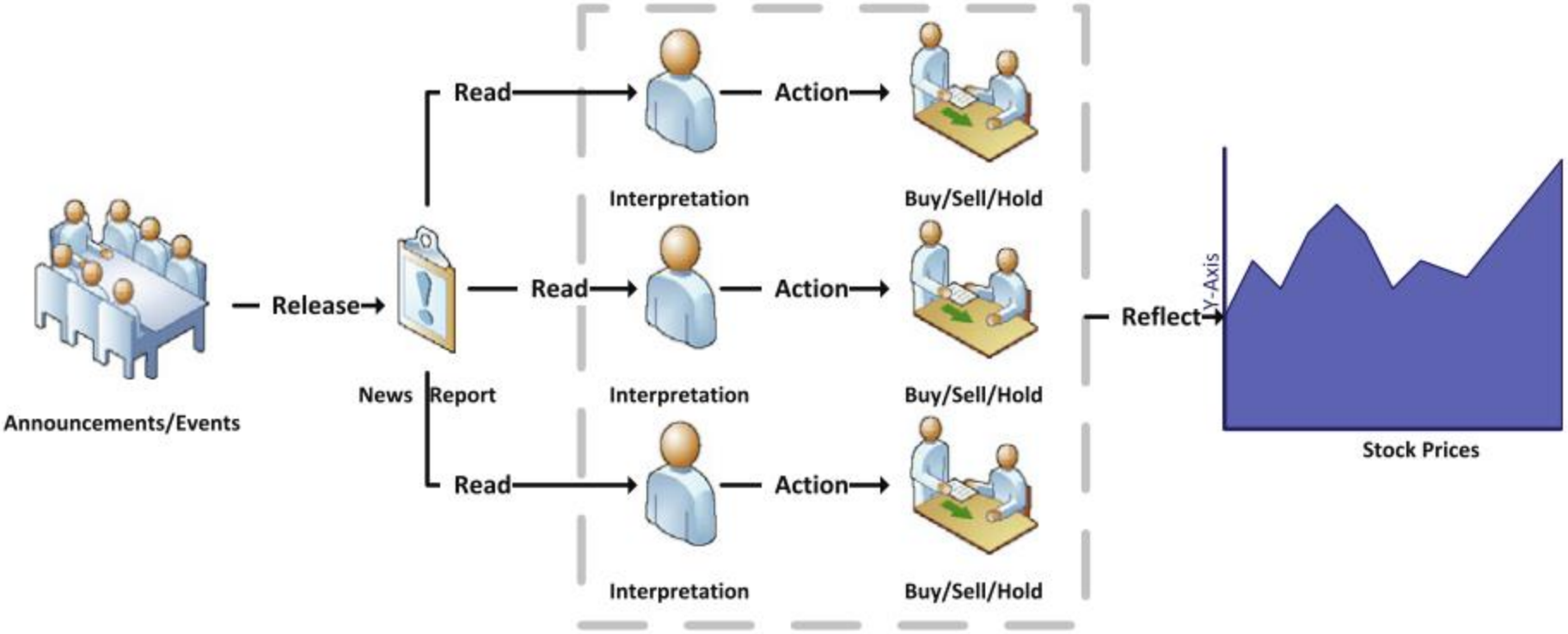
- **Authors**
  - City University of Hong Kong
  - Hong Kong Baptist University（香港浸会大学）
  - Shanghai Jiaotong University
- **Dates**
  - Received 31 October 2013
  - Accepted 15 April 2014
- **Journal**
  - Knowledge-Based Systems
  - Impact Factor: 3.058

Introduction

- **Stock market** is an important and active part of nowadays financial market.
- **Financial news articles** , known as one major source of market information, are widely used and analyzed
- Previous works model news pieces in **bag-of-words space**
- **News sentiment**, which is an important ring on the chain of mapping from the word patterns to the price movements, **is rarely touched.**

Introduction (*cont.*)

- **The general scenario that news impact takes effect on the market prices**



股票指数

- **股票价格指数（stock index）：股票市场总的价格水平变化的指标**
- **道琼斯指数，是算术平均股价指数**
  - 道琼斯工业股价平均指数：30家著名的工业公司
  - 道琼斯运输业股价平均指数：20家著名的交通运输业公司
  - 道琼斯公用事业股价平均指数：15家著名的公用事业公司
  - 道琼斯股价综合平均指数
- **标准•普尔股票价格指数：（标准普尔500指数）标准•普尔公司编制的股票指数**

股票指数

- **恒生指数（Hang Seng Index），香港股市价格的重要指标，指数由若干只成份股市值计算出来**
- **今日恒生指数的计算公式：**

$$CI = \frac{\sum [P(t) \times IS \times FAF \times CF]}{\sum [P(t\text{-}1) \times IS \times FAF \times CF]} \times YCI$$

  - CI：现时指数
  - YCI：上日收市指数
  - P（t）：现时股价
  - P（t-1）：上日收市股价
  - IS：已发行股票数量
  - FAF：流通系数
  - CF：比重上限系数

- **汇丰控股（15%），其次是中国建设银行（7.46%）、中国移动（6.97%）、友邦保险（5.79%）**

Sentiment Dictionary Construction

- **Semi-automatic: seed words + the rules**
  - Adjectives that are separated by ‘and’ have the same polarity
  - Adjectives that are separated by ‘but’ have opposite polarity
  - Synonyms have the same polarity
  - Antonyms have the opposite polarity
- **Manual**
  - Constructed by linguistic experts
  - Accurate
  - Small

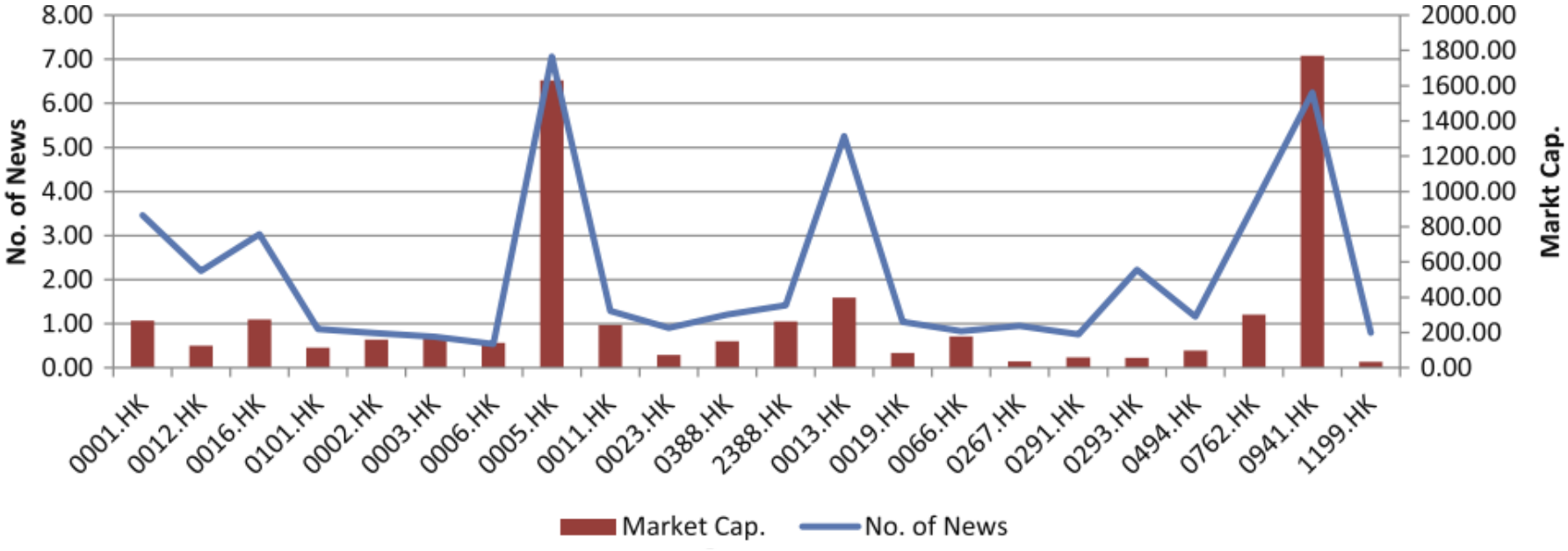
Stock Universe

- **Hong Kong Stock Exchange**
- **Commerce, Finance, Properties and Utilities**
- **Remove the stocks that were added after the starting date (2003-01-01) of our data set**

Sector	Commerce	Finance	Properties	Utilities
Symbol	0013.HK	0005.HK	0001.HK	0002.HK
	0019.HK	0011.HK	~~0004.HK~~	0003.HK
	~~0027.HK~~	0023.HK	0012.HK	0006.HK
	0066.HK	0388.HK	0016.HK	~~0836.HK~~
	~~0135.HK~~	~~0939.HK~~	~~0017.HK~~	
	~~0144.HK~~	~~1299.HK~~	~~0083.HK~~	
	~~0151.HK~~	~~1398.HK~~	0101.HK	
	0267.HK	~~2318.HK~~	~~0688.HK~~	
	0291.HK	2388.HK	~~1109.HK~~	
	0293.HK	~~2628.HK~~		
	~~0322.HK~~	~~3328.HK~~		
	~~0386.HK~~	~~3988.HK~~		
	0494.HK			
	~~0700.HK~~			
	0762.HK			
	~~0857.HK~~			
	~~0883.HK~~			
	0941.HK			
	~~0992.HK~~			
	~~1044.HK~~			
	~~1088.HK~~			
	1199.HK			
	~~1880.HK~~			
	~~1898.HK~~			
	~~1928.HK~~			
Total:	10	5	4	3

Data Set - News

- News archive： FINET（财华网）
- January 2003 to March 2008
- Companies are listed at the end of the news using their stock symbols



Data Set – Stocks Daily Quotes

- **Yahoo! Finance**
- **Open, High, Low, and Close prices**

Symbol	Company Name	Price	Change
GE	General Electric Com...	27.02	0.00
BAC	Bank of America Cor...	15.57	0.00
AAPL	Apple Inc.	127.60	+2.85
MSFT	Microsoft Corporation	42.90	+1.29
PBR	Petróleo Brasileiro S....	8.77	0.00
NOK	Nokia Corporation	7.61	0.00
AMD	Advanced Micro Devi...	2.49	-0.0850
QQQ	PowerShares QQQ T...	107.60	+1.59
FB	Facebook, Inc.	83.09	+2.31
HAL	Halliburton Company	47.85	0.00

Sentiment dictionary

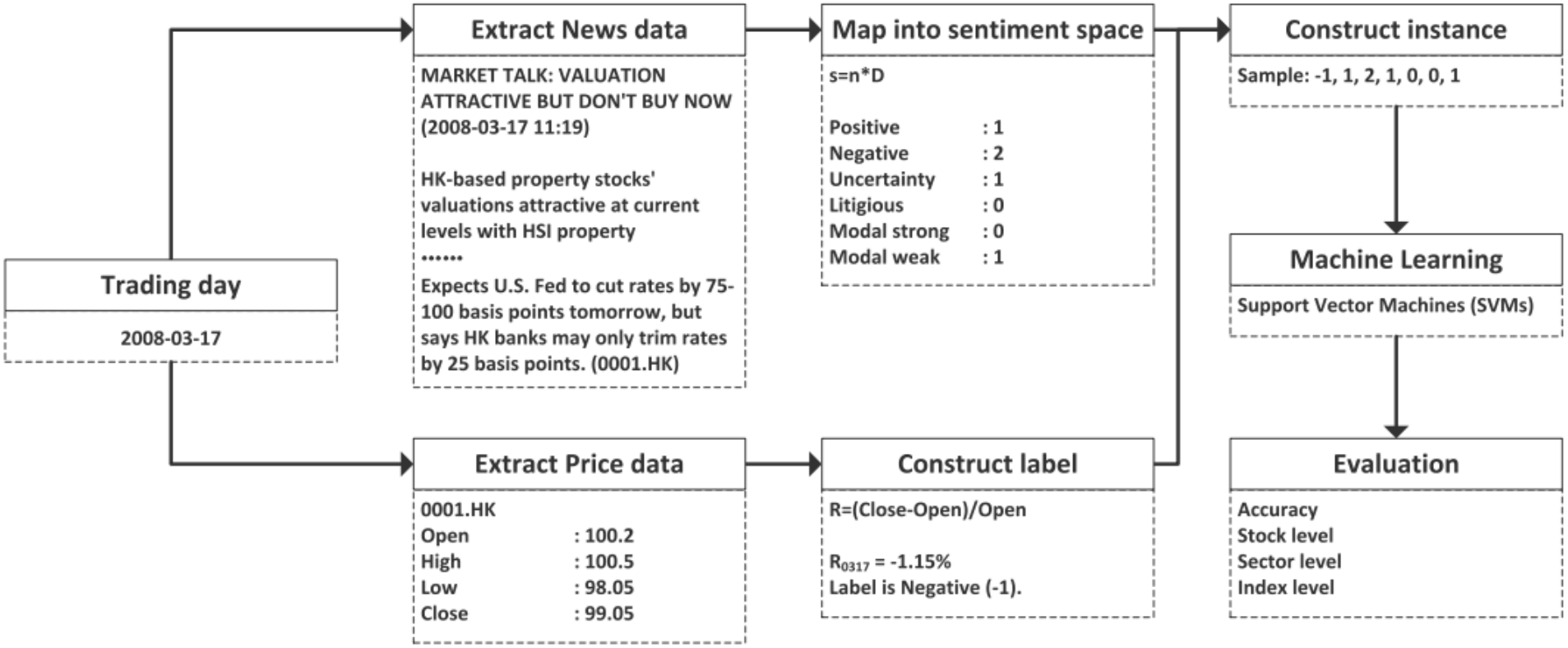
- **Harvard IV-4 sentiment dictionary (HVD)**

No.	Description
1	Positive vs. negative
2	"Osgood" semantic dimensions
3	Pleasure, pain, virtue and vice
4	Overstatement and understatement
5	Language of a particular "institution"

- **Loughran–McDonald financial sentiment dictionary (LMD)**

No.	Description	No. of words
1	Negative words	2349
2	Positive words	354
3	Uncertainty words	291
4	Litigious words	871
5	Modal words strong	19
6	Modal words weak	27

The Generic Framework



Preprocessing of daily news

- **All the news about the same stock are first concatenated as one piece**
- **Translate the daily news into a vector of terms**

$$N = \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_l \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & \dots & t_{1m} \\ t_{21} & t_{22} & \dots & t_{2m} \\ \vdots & & \ddots & \vdots \\ t_{l1} & t_{l2} & \dots & t_{lm} \end{bmatrix}$$

- **Mapping to sentiment space**

$$S = N \odot D$$

Learning and Evaluation

- **Learning**
  - Support Vector Machines (SVMs)
- **Evaluation**

	Predict +	Predict 0	Predict −
True +	t_{++}	f_{+0}	f_{+-}
True 0	f_{0+}	t_{00}	f_{0-}
True −	f_{-+}	f_{-0}	t_{--}

  - accuracy

$$acc = \frac{t_{++} + t_{00} + t_{--}}{all}$$

Preprocessing of daily prices

- **Daily Open-to-Close price return:**

$$R = \frac{Close - Open}{Open}$$

- **Labeling**

$$L(x) = \begin{cases} positive & \text{if } R(x) \geqslant th \\ neutral & \text{if } -th < R(x) < th \\ negative & \text{if } R(x) \leqslant -th \end{cases}$$

A Running Case

- **A piece of news**

A sample piece of news.

MARKET TALK: VALUATION ATTCIVE BUT DON'T BUY NOW (2008–03-17 11:19)
HK-based property stocks' valuations attractive at current levels with HSI property subindex down 4.9% at 2
advises investors not to buy now as sentiment remains uncertain. "The sector is generally trading at sli
attractive," but HK bourse is not going to be immune to more expected selloffs in US markets, he says. Tip
at 21,253.42. Expects US Fed to cut rates by 75–100 basis points tomorrow, but says HK banks may only
HKD 99, NWD (0017.HK) down 8.9% at HKD 15.64 and Sino Land (0083.HK) down 8.7% at HKD 15.82

- **The vector of term frequency value.**

Term	Freq.	Term	Freq.	Term	Freq.
Market	2	Subindex	1	Expect	2
Talk	1	Down	5	Selloff	1
Valuation	2	Advise	1	Tip	1
Attractive	3	Investor	1	Test	1
But	4	Sentiment	1	Key	1

A Running Case

- **News represented by a vector of sentiment value**

	Positive	Negative	Uncertainty	Litigious	Modal strong	Modal weak
Attractive	1	0	0	0	0	0
Uncertain	0	0	1	0	0	1
Short	0	1	0	0	0	0
Cut	0	1	0	0	0	0
s:	1	2	1	0	0	1

A Running Case

- **News represented by a vector of sentiment value**

	Positive	Negative	Uncertainty	Litigious	Modal strong	Modal weak
Attractive	1	0	0	0	0	0
Uncertain	0	0	1	0	0	1
Short	0	1	0	0	0	0
Cut	0	1	0	0	0	0
s:	1	2	1	0	0	1

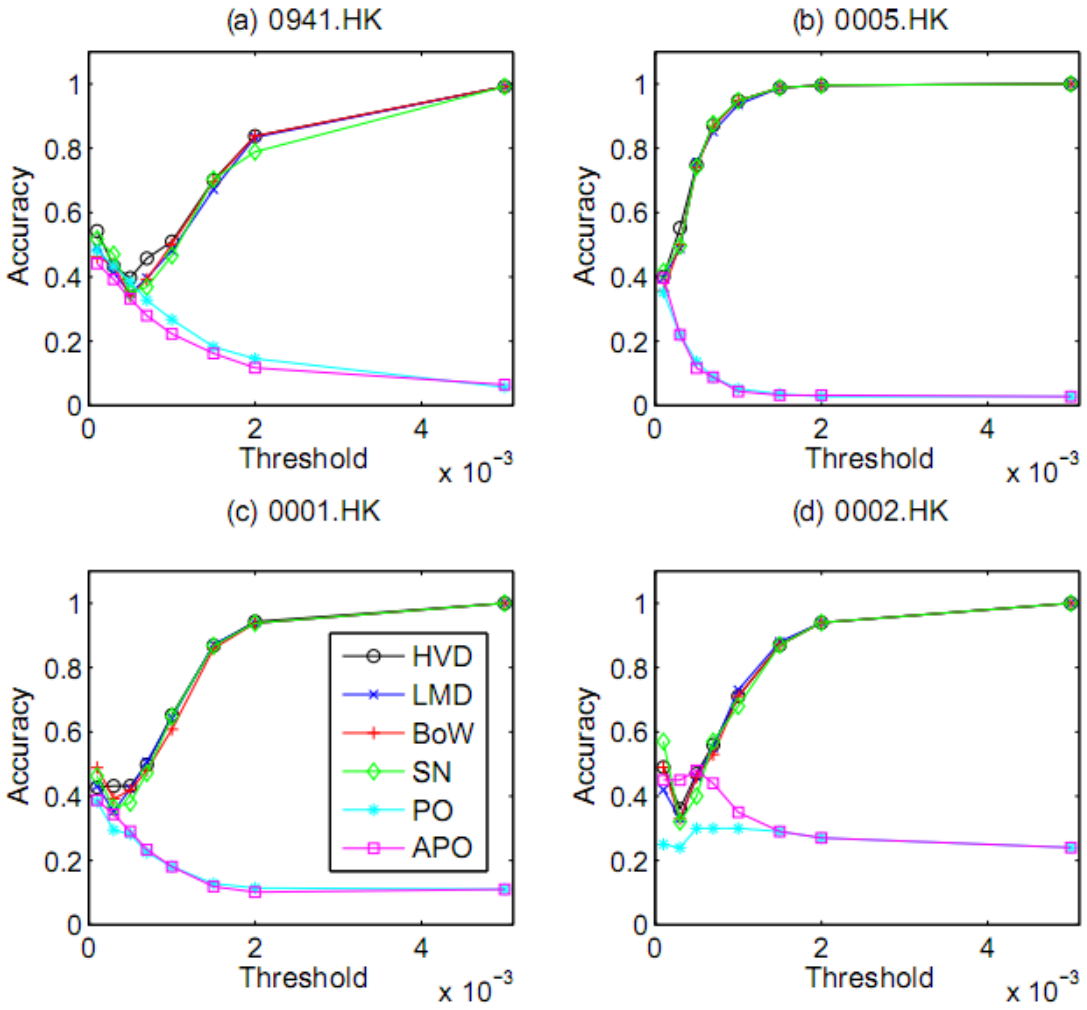
Experiment setup

- **Six different approaches**
  - HVD
  - LMD
  - Bag-of-words (BoW)
  - SenticNet (SN)
  - Sentiment polarity (PO)
  - APO

$$PO = \begin{cases} positive & \text{if } \frac{f(pos)-f(neg)}{f(neg)} \geqslant th \\ negative & \text{if } \frac{f(neg)-f(pos)}{f(pos)} \geqslant th \\ neutral & \text{otherwise} \end{cases}$$

$$APO = \begin{cases} positive & \text{if } \frac{f(neg)-f(pos)}{f(pos)} \geqslant th \\ negative & \text{if } \frac{f(pos)-f(neg)}{f(neg)} \geqslant th \\ neutral & \text{otherwise} \end{cases}$$

Threshold in Labeling Method



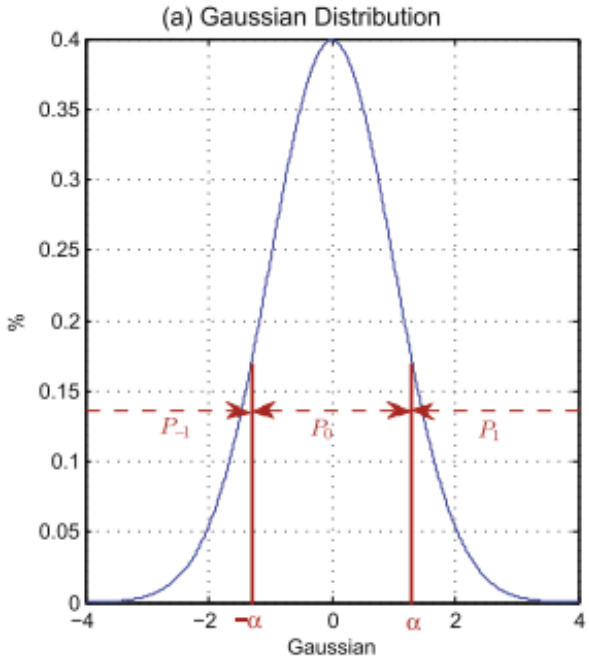
Threshold in Labeling Method

- **Open-to-Close return: Gaussian distribution**

$$P_{-1} = \int_{-\infty}^{-\alpha} \mathrm{pdf}_{Gaussian}(x)\,dx,$$

$$P_{0} = \int_{-\alpha}^{\alpha} \mathrm{pdf}_{Gaussian}(x)\,dx,$$

$$P_{+1} = \int_{\alpha}^{+\infty} \mathrm{pdf}_{Gaussian}(x)\,dx.$$



- **Given the label distribution without extra learning, people can conduct a random draw based on the prior distribution and make predictions.**

Threshold in Labeling Method

- **The accuracy of this approach can be calculated by**

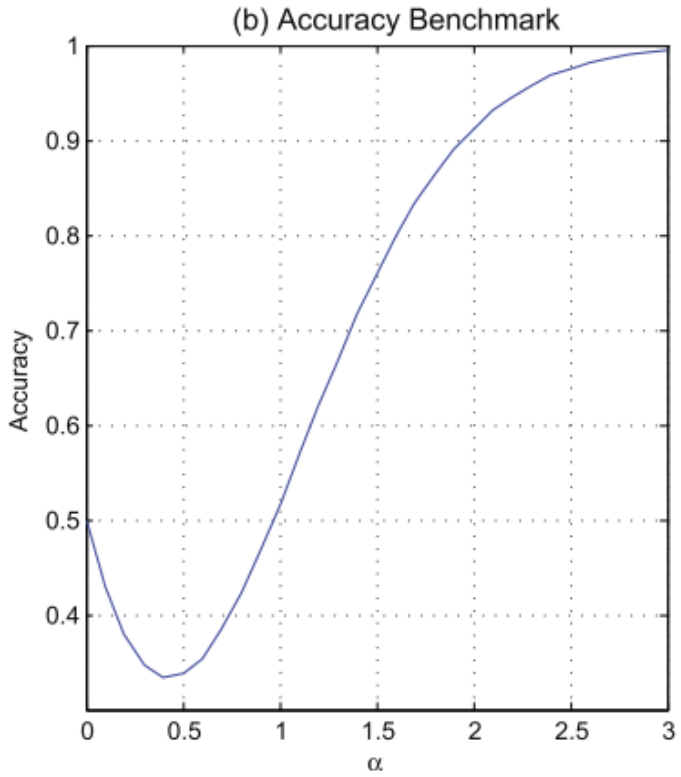
$$acc = P_{-1}^2 + P_0^2 + P_{+1}^2$$

- **With**

$$P_{-1} = P_{+1}$$

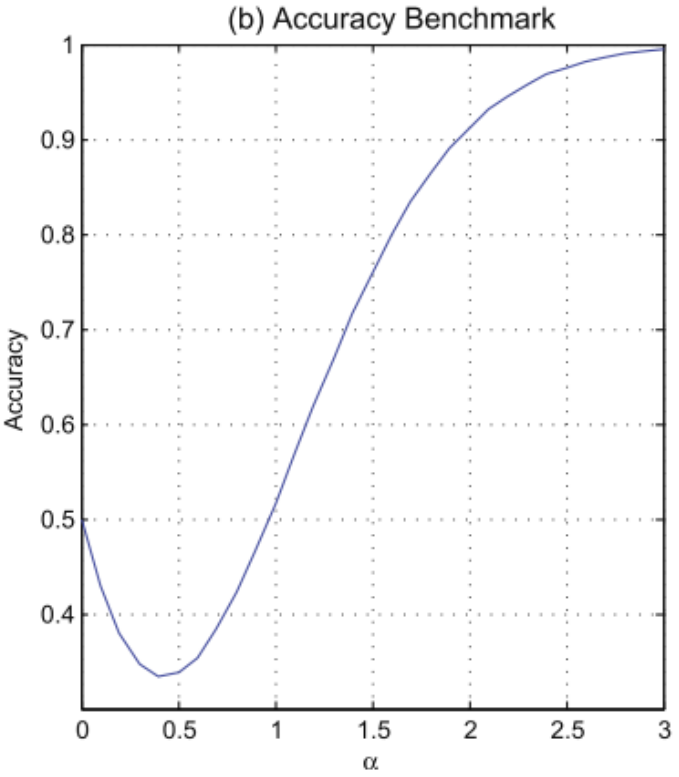
$$P_0 = 1 - 2P_{+1}$$

- **Therefore:** $acc = 6P^2 - 4P + 1$



Threshold in Labeling Method

- The minimum value at $P = \frac{1}{3}$
- When *th* increases, most of the samples are labeled with *neutral*.
- Without learning the prior label distribution, PO and APO are different from others.
- Choose *th* = 0.005 (50 bps)



Individual stock level comparison

Properties						
Symbol	HVD	LMD	BoW	SN	PO	APO
0013.HK	0.3723	0.3628	0.4033	0.3461	0.2816	0.3793
0019.HK	0.4118	0.4902	0.4575	0.3856	0.3793	0.2759
0066.HK	0.4107	0.3839	0.3304	0.4196	0.3818	0.3455
0267.HK	0.4041	0.3630	0.3973	0.3904	0.3889	0.3333
0291.HK	0.3770	0.3525	0.4672	0.4426	0.3500	0.3500
0293.HK	0.3421	0.3872	0.3534	0.3759	0.3109	0.3361
0494.HK	0.3764	0.3258	0.4045	0.4382	0.4048	0.3571
0762.HK	0.3409	0.3750	0.3665	0.3523	0.4380	0.3504
0941.HK	0.3644	0.3583	0.3381	0.3725	0.4170	0.3644
1199.HK	0.4580	0.4733	0.4885	0.4504	0.4800	0.3400
Properties						
Symbol	HVD	LMD	BoW	SN	PO	APO
0001.HK	0.3883	0.3908	0.3714	0.3471	0.3622	0.3351
0012.HK	0.4028	0.3611	0.3715	0.3958	0.3740	0.2977
0016.HK	0.3462	0.3764	0.3874	0.4066	0.4251	0.3713
0101.HK	0.4673	0.4860	0.4766	0.4486	0.4087	0.2870

Individual stock level comparison

Utilities						
Symbol	HVD	LMD	BoW	SN	PO	APO
0005.HK	0.6948	0.6968	0.6767	0.6929	0.2016	0.1895
0011.HK	0.5094	0.5472	0.5236	0.5660	0.3482	0.3929
0023.HK	0.3756	0.3503	0.3807	0.3503	0.3939	0.3030
0388.HK	0.3611	0.3785	0.3472	0.3750	0.3758	0.3576
2388.HK	0.3716	0.3257	0.3716	0.3807	0.3034	0.4045

Utilities						
Symbol	HVD	LMD	BoW	SN	PO	APO
0002.HK	0.4037	0.4224	0.4348	0.3789	0.2459	0.3934
0003.HK	0.3457	0.4198	0.4012	0.3765	0.3056	0.3056
0006.HK	0.2963	0.3796	0.3611	0.3981	0.2558	0.4419

Individual stock level comparison

- **Validation data set**

	HVD	LMD	BoW	SN	PO	APO
HVD	–	11 vs. 11	12 vs. 10	10 vs. 12	20 vs. 2	18 vs. 4
LMD	–	–	14 vs. 8	11 vs. 11	19 vs. 3	19 vs. 3
BoW	–	–	–	12 vs. 10	20 vs. 2	18 vs. 4
SN	–	–	–	–	20 vs. 2	20 vs. 2
PO	–	–	–	–	–	13 vs. 9
APO	–	–	–	–	–	–

Individual stock level comparison

- **Independent testing data set**

	HVD	LMD	BoW	SN	PO	APO
HVD	–	9 vs. 13	8 vs. 14	10 vs. 12	15 vs. 7	17 vs. 5
LMD	–	–	12 vs. 10	12 vs. 10	14 vs. 8	17 vs. 5
BoW	–	–	–	11 vs. 11	14 vs. 8	17 vs. 5
SN	–	–	–	–	15 vs. 7	18 vs. 4
PO	–	–	–	–	–	16 vs. 6
APO	–	–	–	–	–	–

Sector level comparison

- **Sector accuracy**

$$acc_{sector} = acc_{stock} \cdot weight_{sector}.$$

	HVD	LMD	BoW	SN	PO	APO
Commerce	**0.4086**	0.3915	0.3920	0.4000	0.3531	0.3154
Finance	0.6931	**0.7061**	0.6927	0.6976	0.1646	0.1532
Properties	0.4320	**0.4373**	0.4268	0.4063	0.2854	0.2996
Utilities	0.4080	**0.4442**	0.4134	0.4231	0.3147	0.3988

	HVD	LMD	BoW	SN	PO	APO
Commerce	0.3758	0.3763	0.3776	0.3853	**0.3892**	0.3540
Finance	0.6483	**0.6509**	0.6318	0.6408	0.2245	0.2164
Properties	0.3976	**0.4030**	0.3956	0.3858	0.3857	0.3277
Utilities	0.3307	**0.4001**	0.3868	0.3877	0.2707	0.3884

Index Level Comparison

- **Sector Level accuracy**

$$acc_{index} = acc_{stock} \cdot weight_{index}.$$

	HVD	LMD	BoW	SN	PO	APO
Validation	0.5892	**0.5976**	0.5858	0.5876	0.2230	0.2172
Independent testing	0.5460	**0.5527**	0.5391	0.5445	0.2789	0.2665

Conclusion

- **Sentiment analysis does help improve the prediction accuracy**
- **Simply focusing on positive and negative dimensions could not bring useful predictions**
- **There is a minor difference between the models using two different sentiment dictionaries.**

Next Week

- Language Models
- Distributed Representation
- Applications of Distributed Vectors

谢谢！