On the Construction and Analysis of Financial Time-Series-Oriented Lexicons

Cheng-Yi Lai,[1] Chuan-Ju Wang,[1] Ming-Feng Tsai[2]
[1]University of Taipei, Taipei, Taiwan
[2]National Chengchi University, Taipei, Taiwan

June 22, 2015

Outline

1 Introduction

2 Methodology

3 Preliminary Experiments

4 Conclusions and Future work

Outline

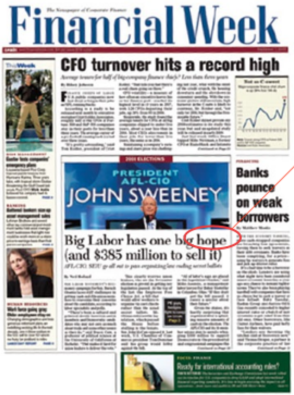
1 Introduction

2 Methodology

3 Preliminary Experiments

4 Conclusions and Future work

Introduction



Financial Week

CFO turnover hits a record high

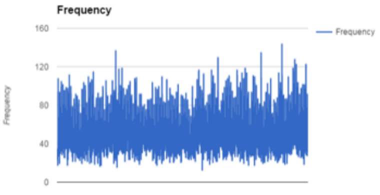
2008 ELECTIONS

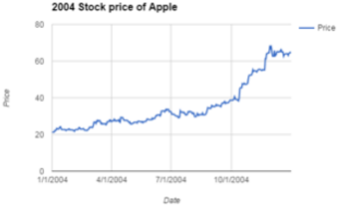
PRESIDENT AFL-CIO JOHN SWEENEY

Big Labor has one big hope (and $385 million to sell it)

Banks pounce on weak borrowers

Ready for international accounting rules?





Introduction

- Text analytics refers to the process of deriving high-quality information from textual information.
- Widely applied to many fields:
  - Biomedicine
  - Finance
  - Social science
- Usage of textual analysis in Finance
  - News articles[1]
  - Financial reports[2]
  - User tweets about publicly-traded companies[3]

[1] Robert and Chen (2009), TOIS

[2] Wang and Tsai (2013), ECIR

[3] Yuexin et.al. (2013), SNAKDD

Introduction

- Sentiment lexicon is a very important resource and can be various in different fields.
- First sentiment lexicon in finance is proposed by Loughran and McDonald (2011) in *Journal of Finance*.
- This lexicon has been widely used in several financial problems.
  - Financial risk prediction[1]
  - Stock movement prediction

[1] Wang and Tsai (2013), ECIR

Introduction

- However, the financial sentiment lexicon has the following limitations.
  1. The lexicon is constructed via only the 10-K financial reports.
     - The wording is formal.
     - Words used in different sources cannot be recognized.
       → E.g., news articles and social networks
  2. The lexicon has no explicit link with the targets of prediction problems.
     - May cause the difficulty in analyzing the obtained prediction models.

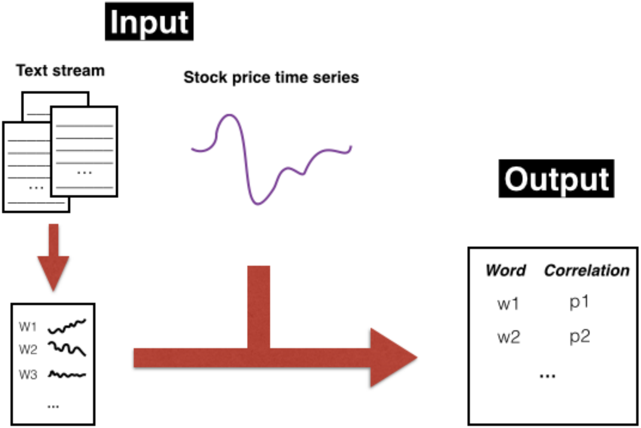
Introduction

- We propose a novel framework to build a time-series-oriented lexicon.
    - Cover different types of sources
    - Have explicit links with the targets of prediction problems
    - Help us build a lexicon to capture more target-oreinted information

Outline

1. Introduction
2. Methodology
3. Preliminary Experiments
4. Conclusions and Future work

Framework



Word	Correlation
w1	p1
w2	p2
...	

Figure : Framework

Methodology

Text stream

1. The text steam: *The New York Times Annotated Corpus*[1] (from 1/1/2004 to 12/31/2006)
2. Use Lemur to index the text stream
   - Stop words, e.g., a ,an, the ,....
   - Stermmer: (Buy , Bougtht , Buying) ▶ Buy
3. Obtain the time series of each word's frequencies

Stcok time series

1. The stock time series: WRDS[2] (from 1/1/2004 to 12/31/2006)
2. Daily stock prices of each company

[1] https://catalog.ldc.upenn.edu/LDC2008T19

[2] https://wrds-web.wharton.upenn.edu/wrds/

Methodology

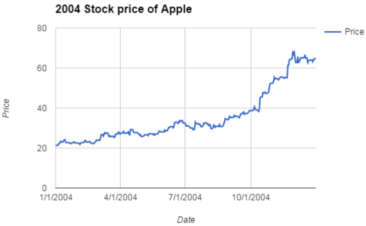


Figure : Stock prices of **Apple**

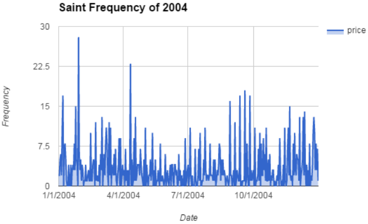


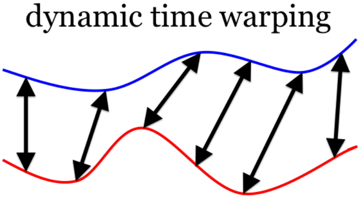
Figure : "Saint" word frequency

- Pearson product-moment correlation coefficient:
  - Stock price of a company: X
  - Word frequency of a certain word: Y
- Pearson Correlation:

$$\rho_{X,Y} = cor(X, Y) = \frac{COV(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Methodology

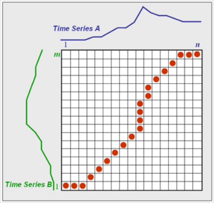
- Low correlation problem:
  - Shifted
  - Stretched
- Solution: *Dynamic Time wraping*



Methodology

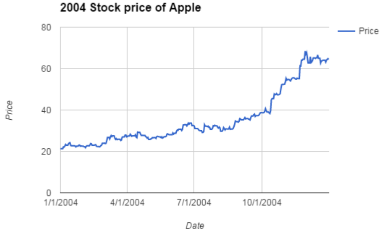
Dynamic Time wraping

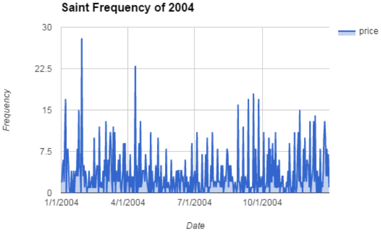
Dynamic time warping (DTW) is an algorithm for measuring similarity between two temporal sequences which may vary in time or speed.



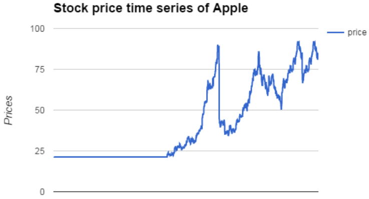
$$\begin{aligned} DTW(X,Y) &= C_p * (X,Y) \\ &= min\{C_p(X,Y), p \in P^{N*M}\} \\ &= D(N,M) \end{aligned}$$

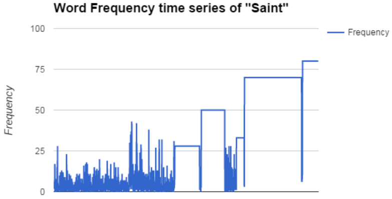
Methodology





Correlation = 0.146451





Correlation = 0.902166

Outline

1 Introduction

2 Methodology

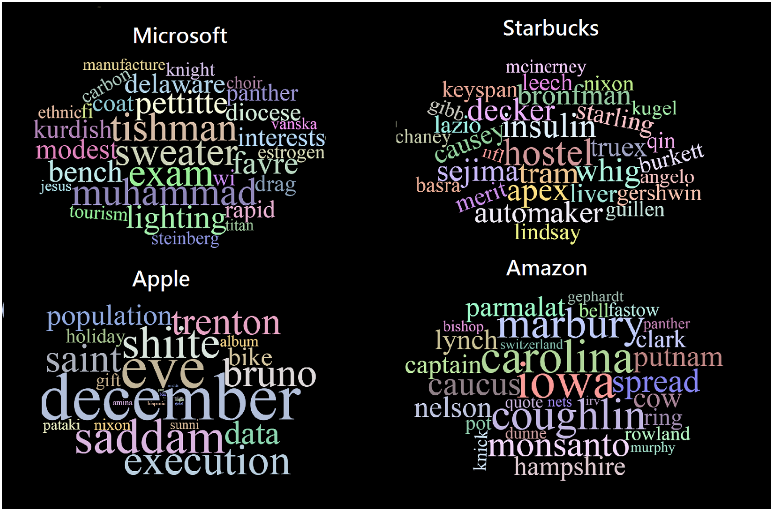
3 Preliminary Experiments

4 Conclusions and Future work

Preliminary Experiments

- Dataset:
  - Text stream: *The New York Times Annotated Corpus*
    - From 1/1/2004 to 12/31/2006
    - #Unique terms: 368509
    - #Documents: 1096
  - Stock time series: WRDS
    - From 1/1/2004 to 12/31/2006
    - Four companies: Apple, Microsoft, Starbucks, Amazon

Preliminary Experiments

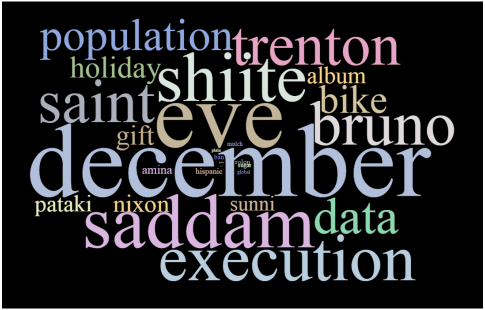

Microsoft

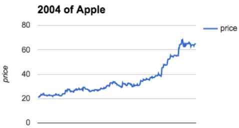
Starbucks

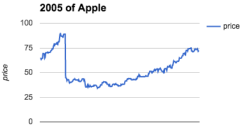
Apple

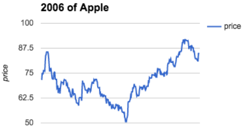
Amazon

Preliminary Experiments









Outline

1 Introduction

2 Methodology

3 Preliminary Experiments

4 Conclusions and Future work

Conclusions and Future work

1. We purpose a framework to construct a target-orient lexicon.
    - It contains all the highly correlation words with target stock prices.
2. For future work, we will validate the resulting lexicons by the task of
    - Predicting stock price rise or fall
    - Predicting financial risk