Enabling Sophisticated Financial Text Mining

Calum Robertson
Research Analyst, Sirca

Overview

• Background
• Data
• Research Strategies
• Obstacles
• Conclusions
Background

• Efficient Market Hypothesis
  – Asset Price reflects all available information

• Sources of Information
  – Trading Data
    • Numerical (i.e., Price, Volume, etc.)
  – News
    • Textual (i.e., Macroeconomic Announcements, Earnings Announcements, Press Announcements)

Background

• Purpose of Market Efficiency Research
  – Academic
    • Evaluate efficiency of market
    • Speculate reasons for supposed inefficiencies (Behavioural Finance)
  – Commercial
    • Exploit market inefficiencies for profit (Algorithmic traders)
    • Manage Risk (Fund Managers)
    • Market Regulation (i.e., prevent insider trading)
Background

• Limitations of Market Efficiency Research
  – Many factors can contribute to the efficiency of a market
    • Some unmeasurable
  – Data
    • Availability
    • Cost
    • Reliability
  – High Frequency Research
    • Noise is common
    • Computationally expensive to identify patterns

Data

• Sources of High Frequency Trading Data
  – Financial Data Providers
    • Thomson Reuters
    • Bloomberg

• Sources of News
  – Newspapers
  – Websites
  – Governments (Macroeconomic Announcements)
  – Financial Data Providers
    • Thomson Reuters
    • Bloomberg
Data

• Sirca
  – Thomson Reuters Integrated Data Network (IDN)
    • 1996 – present
    • Total of over 400 TB of compressed data
      – Currently receive ~ 50 GB of data per day
    • Includes over 130 GB of compressed news
      – Currently 19 supported languages
      – Currently receive up to 200,000 stories per day

Research Strategies

• Event Studies
  – Identify abnormalities in the Time Series and attempt to find Factors which correlate
    • E.g., the effect of Macroeconomic news on Foreign exchange rates
  – Identify Factors suspected to affect Efficiency (e.g., news) and Evaluate the Time Series before and after
    • E.g., the effect of a different types of news on stock prices
Research Strategies

• **Text Classification**
  - Categorise News Based on abnormalities in Time Series
  - Train Classifier based on News Content and Document Category
  - Predict Time Series abnormality after news
  - E.g., Predict 3% returns within 60 minutes of Press Announcement
  - E.g., Predict Volatility of returns with various time windows exceeding several standard deviations from the mean

• **Sentiment Analysis**
  - Analyse the frequency of Positive and Negative terms/phrases in News to Predict Market Movement
  - E.g., Sentiment Analysis of a popular column to evaluate Market Wide movement on following day.
Obstacles

• Data Volume
  – Problem: Too much information for most users
  – Solution:
    • Allow users to acquire aggregated time series of trading data (e.g., minute by minute, hourly, daily)
      – Sirca has successfully provided this service for years
    • Allow users to perform full text search to accurately identify the type of news they require.
      – Sirca has developed services to allow customers to perform searches using any combination of
        » Topic Codes
        » Company Codes
        » Full Text

• Language
  – Problem: News is delivered in 19 different languages
  – Solution: Utilise IBM’s International Components of Unicode (ICU) tool to parse text in any language.

• Text Classification & Sentiment Analysis
  – Problem: Large number of features in the text make it difficult to perform large scale text classification and sentiment analysis
  – Solution:
    • Calculate Term Counts for every news article
    • Produce Report of term frequencies within documents matching user search criteria
Obstacles

• Usability
  – Problem: Many complex, though routine, operations required to perform sophisticated financial text mining.
  – Solution: Developing a suite of Web Services to allow users to easily combine our data and standard techniques to conduct their own novel research.

Conclusions

• Many Obstacles to Large Scale Financial Text Mining
• Sirca is actively working to promote research within the field.
Questions