Text and Context: Language Analytics in Finance

Sanjiv Ranjan Das
Santa Clara University
Leavey School of Business
srdas@scu.edu
Foundations and Trends® in Finance
Volume 8, Issue 3, 2014
Editorial Board

Editor-in-Chief

George M. Constantinides
Booth School of Business
University of Chicago
United States

Editors

Richard Green
Co-Editor
Carnegie Mellon University
Francis Longstaff
Co-Editor
University of California, Los Angeles
Sheridan Titman
Co-Editor
University of Texas at Austin
Editorial Scope

Topics

Foundations and Trends® in Finance publishes survey and tutorial articles in the following topics:

- Corporate finance
  - Corporate governance
  - Corporate financing
  - Dividend policy and capital structure
  - Corporate control
  - Investment policy
  - Agency theory and information

- Financial markets
  - Market microstructure
  - Portfolio theory
  - Financial intermediation
  - Investment banking
  - Market efficiency
  - Security issuance
  - Anomalies and behavioral finance

- Asset pricing
  - Asset-pricing theory
  - Asset-pricing models
  - Tax effects
  - Liquidity
  - Equity risk premium
  - Pricing models and volatility
  - Fixed income securities

- Derivatives
  - Computational finance
  - Futures markets and hedging
  - Financial engineering
  - Interest rate derivatives
  - Credit derivatives
  - Financial econometrics
  - Estimating volatilities and correlations

Information for Librarians

Foundations and Trends® in Finance, 2014, Volume 8, 4 issues. ISSN paper version 1567-2395. ISSN online version 1567-2409. Also available as a combined paper and online subscription.
Text and Context:
Language Analytics in Finance

Sanjiv Ranjan Das
Santa Clara University
Leavey School of Business
srdas@scu.edu
4.5 Vector distance classifiers .......................... 52

5 Metrics .......................... 54
  5.1 Confusion Matrix .................................. 55
  5.2 Accuracy ........................................... 57
  5.3 False Positives ..................................... 58
  5.4 Sentiment Error ..................................... 59
  5.5 Disagreement ....................................... 60
  5.6 Correlations ......................................... 60
  5.7 Phase lag metrics .................................... 61
  5.8 Readability ......................................... 63

6 Applications and Empirics .................. 67
  6.1 Predicting Market Movement ...................... 69
  6.2 Predicting risk, volatility, volume ................ 78
  6.3 Text Mining Company Reports ..................... 79
  6.4 Text Mining Public Data and Network Modeling .... 84
  6.5 News Analytics ..................................... 89
  6.6 Commercial Vendors ................................ 94

7 Text Analytics – The Future .................. 100

Acknowledgements .................................. 106

Appendices ............................................ 107

A Sample text from Bloomberg for summarization .... 108

References ............................................. 113
Abstract

This monograph surveys the technology and empirics of text analytics in finance. I present various tools of information extraction and basic text analytics. I survey a range of techniques of classification and predictive analytics, and metrics used to assess the performance of text analytics algorithms. I then review the literature on text mining and predictive analytics in finance, and its connection to networks, covering a wide range of text sources such as blogs, news, web posts, corporate filings, etc. I end with textual content presenting forecasts and predictions about future directions.
What is Text Mining?

Howard: You know, I’m really glad you decided to learn Mandarin.
Sheldon: Why?
Howard: Once you’re fluent, you’ll have a billion more people to annoy instead of me.

"The Tangerine Factor"
The Big Bang Theory, Season 1, Episode 17

If you consider all the data in the universe, only some of it is in numerical form. There is certainly a lot more text.¹ If you read a financial news article, the quantity of text vastly outnumbers the quantity of numbers. Until recently, financial analysis was just based on numbers. Usage of text required human coding of attributes into numerical form before yielding to analysis. This was a slow process, and not exhaustive, given how much textual data is at hand. We are entering the age of

---

¹We may also consider images, sound clips, and videos as data, in which case, numerical data comprises a very small portion of human expression and experience. See Mayew and Venkatachalam [2012] for the use of speech analysis in deciphering the emotive content of voice communications by managers of firms.
Big Text, and this monograph describes the current landscape of text analytics.

Text is versatile. It contains nuances and behavioral expression that is not possible to convey using numbers. Behavioral economics makes a case for considering these nuances that permeate human activity, in economics and finance. Advances in computer science have made text mining possible, and finance is replete with applications, and offers substantial payoffs for profit-making ideas using text mining tools.

There are several benefits to enhancing quantitative financial analysis with text mining analytics. First, text contains emotive content that may be useful in assessing sentiment in markets. There are several articles in mainstream journals that deal with this topic, both theoretical and empirical [for example, Admati and Pfleiderer, 2001, DeMarzo et al., 2003, Antweiler and Frank, 2004, 2005, Das and Chen, 2007, Tetlock, 2007, Tetlock et al., 2008, Mitra et al., 2008, Leinweber and Sisk, 2010].

Second, text contains opinions and connections that may be harvested and assessed for trading rules, or to corroborate other news, or for risk assessment. Many papers examine these issues as well, and present the benefits of such analysis, as in Das et al. [2005], Das and Sisk [2005], Godes et al. [2005], Li [2006], Hochberg et al. [2007].

Third, many facts do not lend themselves to quantitative expression. They may be intrinsically qualitative and better expressed in the form of text. Of course, most qualitative phenomena may be expressed as numerical quantities on a discrete support, but such abstraction results in a loss of holistic meaning. For example, a trading algorithm may examine a news report to determine a buy or sell signal, and text mining tools can use past data on news and trading outcomes to determine the best course of action in a seamless, efficient manner. Coding text using quantitative variables, i.e., dummy variables for the various attributes of text is chunky, spawns too many variables, and is less accurate.

Fourth, numbers tend to aggregate and summarize underlying phenomena, of infinite variety, and the nuances are better expressed in text, which is disaggregated. Numbers are not raw, original data, but
What is Text Mining?

quantifications of characteristics of markets, often first expressed in textual form. For this reason, it is likely that text (such as news streams) contains information that is more timely than numerical financial information, and better suited to predictive analytics. There is evidence that textual information may be used to predict markets, as in Antweiler and Frank [2004], Tetlock [2007], Leinweber and Sisk [2010]. Analyzing large bodies of text enables operationalization of the wisdom of the crowds as discussed in the excellent book by Surowiecki [2004].

The benefits of text mining are easy to see without defining it formally, but it’s time to attempt a formal definition. **Text mining is the large-scale, automated processing of plain text language in digital form to extract data that is converted into useful quantitative or qualitative information.** Hence, text mining is automated on big data that is not amenable to human processing within reasonable time frames. It entails extracting data that is converted into information of many types. Text mining may be simple as in key word searches and counts. Or it may require language parsing and complex rules for information extraction. It may be applied to structured text, such as the information in forms and some kinds of web pages, or it may be applied to unstructured text, a much harder endeavor. Text mining is also aimed at unearthing unseen relationships in unstructured text as in meta analyses of research papers, see Van Noorden [2012].

A subfield of text mining is “news analytics.” Wikipedia defines it as - “…the measurement of the various qualitative and quantitative attributes of textual (unstructured data) news stories. Some of these attributes are: sentiment, relevance, and novelty. Expressing news stories as numbers permits the manipulation of everyday information in a mathematical and statistical way. News analytics are used in financial modeling, particularly in quantitative and algorithmic trading. Further, news analytics can be used to plot and characterize firm behaviors over time and thus yield important strategic insights about rival firms. News analytics are usually derived through automated text analysis and ap-

---

2See the article by Gary Belsky, “Why Text Mining may be The Next Big Thing” in TIME: http://business.time.com/2012/03/20/why-text-mining-may-be-the-next-big-thing/print/.
plied to digital texts using elements from natural language processing and machine learning such as latent semantic analysis, support vector machines, ‘bag of words’, among other techniques.”

In the ensuing chapters we will examine several topics in financial text mining. In Chapter 2 we examine how text is extracted from various web sites and services. Chapter 3 deals with the basics of text analytics such as dictionaries, lexicons, mood scoring, and summarization of text. This is followed by the analytics of text classification in Chapter 4. The performance of text analytic algorithms is assessed using a range of metrics in Chapter 5. A survey of the empirical literature on text mining in finance and the commercialization of textual analytics is discussed in Chapter 6. Finally, we end with a look at the future of text analytics in Chapter 7.
References


References


References


References


