# Text and Context: Language Analytics in Finance

**Sanjiv Ranjan Das**
Santa Clara University
Leavey School of Business
srdas@scu.edu

the essence of knowledge
Boston — Delft

**Foundations and Trends® in Finance**

*Published, sold and distributed by:*
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

*Outside North America:*
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

S. R. Das. *Text and Context: Language Analytics in Finance.* Foundations and Trends® in Finance, vol. 8, no. 3, pp. 145–261, 2014.

*This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.*

ISBN: 978-1-60198-911-6

**Foundations and Trends® in Finance**
Volume 8, Issue 3, 2014

# 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.

Foundations and Trends® in Finance
Vol. 8, No. 3 (2014) 145–261

DOI: 10.1561/0500000045

the essence of knowledge

# Text and Context: Language Analytics in Finance

Sanjiv Ranjan Das
Santa Clara University
Leavey School of Business
srdas@scu.edu

# Contents

## 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.



S. R. Das. *Text and Context: Language Analytics in Finance*. Foundations and Trends® in Finance, vol. 8, no. 3, pp. 145–261, 2014.
DOI: 10.1561/0500000045.

# 1

## 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

[1] We 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 clunky, 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 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].[2]

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-

[2]See 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

A. Admati and P. Pfleiderer. Noisytalk.com: Broadcasting opinions in a noisy environment. Working paper, Stanford University, 2001.

W. Antweiler and M. Frank. Is all that talk just noise? the information content of internet stock message boards. *Journal of Finance*, v59(3):1259–1295, 2004.

W. Antweiler and M. Frank. The market impact of corporate news stories. Working paper, University of British Columbia, 2005.

Mark Bagnoli, M. Beneish, and S. Watts. Whisper forecasts of quarterly earnings per share. *Journal of Accounting and Economics*, 28(1):27–50, 1999.

R. Bar-Haim, E. Dinur, R. Feldman, Fresko M, and G. Goldstein. Identifying and following experts in stock microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1310–1319. Edinburgh, UK, 2011.

A. Bodnaruk, T.Loughran, and B. McDonald. Using 10-k text to gauge financial constraints. Working paper, University of Notre Dame, 2013.

J. Bollen, H. Mao, and X-J. Zeng. Twitter mood predicts the stock market. *arXiv:1010.3003v1*, 2010.

P. Bonacich. Technique for analyzing overlapping memberships. *Sociological Methodology*, 4:176–185, 1972.

P. Bonacich. Power and centrality: A family of measures. *American Journal of Sociology*, 92(5):1170–1182, 1987.

J. Boudoukh, R. Feldman, S. Kogan, and M. Richardson. Which news moves stock prices? a textual analysis. Working paper, University of Texas, Austin, 2012.

M. Bradley and P. Lang. Affective norms for english words (anew): Stimuli, instruction manual and affective ratings. Technical report C-1, *The Center for Research in Psychophysiology*, University of Florida, 1999.

E. D. Brown. Will twitter make you a better investor? a look at sentiment, user reputation and their effect on the stock market. In *Proceedings of the Southern Association for Information Systems Conference*. Atlanta, GA, USA, March 23rd–24th 2012.

D. Burdick, S. Das, M. A. Hernandez, H. Ho, G. Koutrika, R. Krishnamurthy, L. Popa, I. Stanoi, and S. Vaithyanathan. Extracting, linking and integrating data from public sources: A financial case study. *IEEE Data Engineering Bulletin*, 34(3):60–67, 2011.

M. Coleman and T. L. Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284, 1975.

D. Conway and J. M. White. *Machine Learning for Hackers*. O'Reilly Press, Sebastopol, CA, 2012.

S. Das and M. Chen. Yahoo for amazon! sentiment extraction from small talk on the web. *Management Science*, 53:1375–1388, 2007.

S. Das and J. Sisk. Financial communities. *Journal of Portfolio Management*, 31(4):112–123, 2005.

S. Das, A. Martinez-Jerez, and P. Tufano. einformation: A clinical study of investor discussion and sentiment. *Financial Management*, 34(5):103–137, 2005.

S. R. Das. News analytics: Framework, techniques and metrics. In *The Handbook of News Analytics in Finance*. John Wiley & Sons, U.K, 2011.

G. De Franco, O.-K. Hope, D. Vyas, and Y. Zhou. Analyst report readability. *Contemporary Accounting Research*, forthcoming, 2013.

P. DeMarzo, D. Vayanos, and J. Zwiebel. Persuasion bias, social influence, and uni-dimensional opinions. *Quarterly Journal of Economics*, 118:909–968, 2003.

R. Feldman, B. Rosenfeld, R. Bar-Haim, and M. Fresko. The stock sonar sentiment analysis of stocks based on a hybrid approach. *Proceedings of the Twenty-Third Innovative Applications of Artificial Intelligence Conference*, pages 1642–1647, 2011.

C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

D. Godes, D. Mayzlin, Y. Chen, S. Das, C. Dellarocas, B. Pfeieffer, B. Libai, S. Sen, M. Shi, and P. Verlegh. The firm's management of social interactions. *Marketing Letters*, v16:415–428, 2005.

R. Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952.

Y. Hochberg, A. Ljungqvist, and Y. Lu. Whom you know matters: Venture capital networks and investment performance. *Journal of Finance*, 62(1): 251–301, 2007.

P. Jaccard. Etude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

N. Jegadeesh and D. Wu. Word power: A new approach for content analysis. *Journal of Financial Economics*, 110(3):712–729, 2013.

R. Jordan. *Academic Writing Course*. Longman, London, 1999.

J. Lanier. *Who Owns the Future?* Simon and Schuster, New York, 2013.

R. Lehavy, F. Li, and K. Merkley. The effect of annual report readability on analyst following and the properties of their earnings forecasts. *Accounting Review*, 86:1087–1115, 2011.

D. Leinweber. *Nerds on Wall Street*. John Wiley and Sons, New Jersey, 2009.

D. Leinweber and J. Sisk. *Relating News Analytics to Stock Returns*. mimeo, Leinweber & Co, 2010.

D. Leinweber and J. Sisk. Event-driven trading and the "new news". *Journal of Portfolio Management*, Summer, 1–15 2011.

F. Li. Do stock market investors understand the risksentiment of corporate annual reports? Working paper, University of Michigan, 2006.

F. Li. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45:221–247, 2008.

A. Logunov. A tweet in time: Can twitter sentiment analysis improve economic indicator estimation and predict market returns? Undergraduate Honors Thesis, University of New South Wales, 2011.

T. Loughran and W. McDonald. When is a liability not a liability. *Journal of Finance*, 66:35–65, 2011.

T. Loughran and W. McDonald. Measuring readability in financial disclosures. *Journal of Finance*, 69:1643–1671, 2014.

H.-M. Lu, H. Chen, T.-J. Chen, M.-W. Hung, and S.-H. Li. Financial text mining: Supporting decision making using web 2.0 content. *IEEE Intelligent Systems*, pages 78–82, 2010.

C.D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

W. J. Mayew and M. Venkatachalam. Speech analysis in financial markets. *Foundations and Trends in Accounting*, 7(2):73–130, 2012.

A. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/∼mccallum/bow, 1996.

D. McNair, J. P. Heuchert, and E. Shilony. *Profile of Mood States. Bibliography 1964–2002*. Multi-Health Systems, 2003.

G. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

L. Mitra, G. Mitra, and D. diBartolomeo. Equity portfolio risk (volatility) estimation using market information and sentiment. Working paper, Brunel University, 2008.

N. Pervin, F. Fang, A. Datta, and K. Dutta. Fast, scalable, and context-sensitive detection of trending topics in microblog post streams. *ACM Transactions on Management Information Systems*, 3(4):Article 19, 2013.

M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

T. Rao and S. Srivastava. Twitter sentiment analysis: How to hedge your bets in the stock markets. Working paper, Indian Institute of Technology, Delhi, 2012.

R. Roll. R-squared. *Journal of Finance*, 43:541–566, 1988.

T. Sprenger. Tweettrader.net: Leveraging crowd wisdom in a stock micro blogging forum. *Association for the Advancement of Artificial Intelligence*, 2011.

T. Sprenger and I. M. Welpe. Tweets and trades: The information content of stock microblogs. Working paper, Technische Universität München, 2010.

J. Surowiecki. *The Wisdom of Crowds*. Anchor Books, New York, 2004.

P. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3):1139–1168, 2007.

P. Tetlock, P. M. Saar-Tsechansky, and S. Macskassay. More than words: Quantifying language to measure firm's fundamentals. *Journal of Finance*, 63(3):1437–1467, 2008.

R. Tumarkin and R.Whitelaw. News or noise? internet postings and stock prices. *Financial Analysts Journal*, 57(3):41–51, 2001.

R. Van Noorden. Trouble at the text mine. *Nature*, 483:134–135, 2012.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

V. Vapnik and Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, v16(2):264–280, 1964.

V. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and Remote Control*, v24, 1963.

T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Opinionfinder: A system for subjectivity analysis. *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, pages 34–35, 2005.

P. Wysocki. Cheap talk on the web: The determinants of postings on stock message boards. Working Paper, November, University of Michigan, 1999.

W. Zhang and S. Skiena. Trading strategies to exploit blog and news sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 375–378, 2010.