WILEY FINANCE

The Handbook of News analytics in finance

Edited by
GAUTAM MITRA
LEELA MITRA

This edition first published in 2011

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For other titles in the Wiley Finance Series please see www.wiley.com/finance

ISBN 978-0-470-66679-1 (hardback)
ISBN 978-1-119-99080-2 (ebook)
ISBN 978-1-119-97796-4 (ebook)
ISBN 978-1-119-97797-1 (ebook)

A catalogue record for this book is available from the British Library.

Project management by OPS Ltd, Gt Yarmouth, Norfolk
Typeset in 10/12pt Times
Printed in Great Britain by CPI Antony Rowe, Chippenham, Wiltshire

Preface

The purpose of a preface in our view is rather unashamedly to sell the book—to communicate the message of the book succinctly and either to motivate the reader to explore its content or to leave the reader feeling that just maybe he or she is losing out if the book's theme does not fire their imagination. So, by ignoring this book you will never know whether you might have seen the light and gleaned the winning strategies of financial analytics! The subheadings in this preface are deliberately linked to coax you to send an email to your quant team instructing them to pick up this handbook and mine it for nuggets of knowledge. You may also post a review in your blog or alert your peers in Linked-in depending on how much enthusiasm we have been able to generate.

The *background* sets the scene. We then highlight the *research problems* that also equate with the *business problems*. We discuss the *role of news* followed by an outline of the different *technologies* that underpin news analytics (NA). We then emphasize that discovering what the experts—that is, our enthusiastic contributors—have to say can be rewarding. We conclude the preface with a suggested reading strategy—a *road map*—with a view to help the reader make the most of effective knowledge mining.

Background: the setting

Our research base, the Centre for the Analysis of Risk and Optimisation Modelling Applications (CARISMA) was established in 2001 within Brunel University. CARISMA conferences bring together practitioners, hard-nosed business people, and academics, the abstract thinkers. Sometimes this formula works and the academics are puzzled, challenged, and fascinated by the prospect of analyzing a difficult business problem that can also be construed as a research problem. There are many different constituents that make up the financial (news analytics) market place: academics, industry-based quant researchers, news sentiment data vendors and, finally, traders and investment strategy managers. All these people are variously attracted by the prospect of determining the quantified sentiment of the market by analysis of the news. There is one common aspect which brings the contributors of this volume together: namely, they are people with a "can do" spirit who believe with unwavering conviction that they will find the silver bullet.

The research problem = the business problem

The world of financial analytics is concerned with three leading problems:

(i) The pricing of assets in a temporal setting.
(ii) Making optimum investment decisions low frequency or optimum trading decisions high frequency.
(iii) Controlling risk at different time exposures.

The role of news

News provides information about an event and, as such, may be considered to be an event in itself—news moves the market. The dynamics of the flow of information and market uncertainty impacts security price formation, price discovery, market participant behaviour such as price (over) reaction, price volatility, and market stability. Traders and other market participants digest news rapidly; they may revise and rebalance their asset positions. Most traders have access to newswires at their desks. The sources and the volume of news continue to grow.

The technologies underpinning NA

It is widely recognized that news plays a key role in financial markets. New technologies that enable automatic or semi-automatic news collection, extraction, aggregation, and categorization are emerging. Machine-learning techniques are used to process the textual narrative of news stories, thus transforming qualitative descriptions into quantified news sentiment scores. A range of computational models (algorithms) have been proposed for this purpose. Typically, positive-word or negative-word counts or vector distance computation, adjective or adverb phrase usage or the Bayesian approach of introducing domain experts' subjective and contextual knowledge are applied to calculate a sentiment score. In the context of trading, news sentiment data have to be fused with the market data of "trades and quotes" to create an analytic data mart for financial models. Herein lies the challenge of automation. Not only do systems that support information flow have to be designed, they have to be connected to models of financial analytics for asset pricing, trading, investment management, and risk control. Thus, financial engineering goes hand in hand with information engineering to create winning strategies.

The road map

As editors we set the scene in Chapter 1 of the book. In this chapter we provide a general review of applications of NA in finance. We discuss news data sources, methods of turning qualitative text to quantified metrics and a range of models and applications. In particular, we would like to draw the attention of the reader to the two sections of the appendix where we describe in summary form the structure and content of news data as supplied by Thomson Reuters in its News Scope and RavenPack in its News Scores products. The major themes of this handbook are:

Part 1 The methods and models by which news sentiment is measured and quantified.
Part 2 News and abnormal returns as found in trading models and investment strategies.
Part 3 How news analytics can be used for risk control.
Part 4 The insight of industry leaders and relevant commercial information.

Depending on what interests them most, readers may turn their attention to any of these parts, scan the titles and abstracts, and read the articles as they are presented. There is very little interdependence between these four parts of the handbook.

The contributors are either researchers from academia or practitioners from industry—in some instances, both. They have two things in common: they are all experts in NA and they are enthusiastic about applying NA to finance. As editors we believe our salient achievement has been to solicit and convince this team of enthusiasts to contribute their knowledge and their recent research results to this volume. Finally, we would invite readers to contemplate, innovate and be excited by the infectious enthusiasm of the contributors—you may come up with your own rewarding applications of news analytics and hopefully share them with other experts in the field.

Gautam Mitra and Leela Mitra
London

Acknowledgements

Leela and I would like to thank Leela's mother and my dear wife Dhira for her help in putting this volume together. Dhira has helped us in many aspects of editing this book—communicating with the contributors, the publishers, and the sponsors. She has done so always with a smile and she only frowned whenever one of us (Gautam) kept missing the schedule. Without her help we would have missed the boat. We would not have studied and researched news analytics (NA) had we not been invited to spend a brainstorming weekend in early January 2008 at RavenPack's R&D villa in Marbella, Spain. We got smitten by the research challenges that were presented to us; subsequently, one of us, that is, Leela delved deeper into the subject as part of her PhD research. We also realized that NA, despite being in the early stages of its development, holds great promise as a modeling tool to enhance financial analytics. We therefore decided that the information and research results that we are still gathering should be shared widely with practitioners and the academic community by compiling this handbook. The handbook has also been championed by RavenPack and Thomson Reuters. They have contributed financially (platinum sponsors) and have actively solicited on our behalf contributions from industry leaders. Grateful thanks are therefore due to Armando Gonzales and Richard Brown of RavenPack and Thomson Reuters, respectively. The sponsorships of Media Sentiment and Northfield Information Services are also acknowledged. We would also like to thank all the contributors for enthusiastically sharing their research results. OptiRisk organized workshops and forums on NA in 2009 and in 2010; and a number of colleagues promoted, organized, and hosted these events. We would like to record our appreciation to this terrific team comprising Julie Valentine, Michael and Hetty Sun, Chanakya Mitra, and Natallia Zverovich; these events played a spiritually uplifting key role in the compilation of this handbook.

About the editors

Gautam Mitra (London) is an internationally renowned research scientist in the field of computational optimization and modeling. He has developed a world-class research group in his area of specialization with researchers from Europe, the U.K., the U.S.A., and Asia. He has published three books and over one hundred refereed research articles. He was Head of the Department of Mathematical Sciences, Brunel University between 1990 and 2001. In 2001 he established CARISMA: The Centre for the Analysis of Risk and Optimisation Modelling Applications. CARISMA specializes in the research of risk and optimization and their combined paradigm in decision modeling. Professor Mitra is also a Director of UNICOM Seminars and OptiRisk Systems; OptiRisk specializes in the research and development of optimization and financial analytics tools.

Leela Mitra (London) is a quantitative analyst at OptiRisk Systems. Dr Mitra joined OptiRisk Systems in that capacity in 2004. She received her PhD in operational research on the topic of "Scenario generation for asset allocation models" from CARISMA, Brunel University. Topics included: "mixed" scenario sets for investment decisions with downside risk; pricing and evaluating a bond portfolio using a regime-switching Markov model; and desirable properties for scenario generation. She has a first-class BA (joint honours) degree in mathematics and philosophy from King's College (University of London). Prior to joining OptiRisk, Leela worked in the pensions industry as an actuarial consultant for Mercer HR and, subsequently, with Jardine Lloyd Thomson. She is part-qualified as an actuary.

About the contributors

Brad Barber is the Gallagher Professor of Finance at the UC Davis Graduate School of Management where he teaches introductory finance to MBA students. His research focuses on the psychology of individual investors, is widely published in leading academic journals and is frequently referenced in the financial press.

Gurvinder Brar heads the European Quantitative Research Team at Macquarie Securities. He focuses on multifactor stock selection models, style research and small-cap quant strategy. Prior to Macquarie he worked for 8 years at Citi as part of the #1-ranked European Quantitative Research Team. Prior to that Gurvinder spent 2 years in the Risk-adjusted Portfolio Analysis Team for NatWest.

Richard Brown is the Global Business Manager for the Machine Readable News program at Thomson Reuters, responsible for the product portfolio that includes its archive product, real-time feeds, and news analysis solutions.

Sanjiv Das is Professor of Finance at Santa Clara University. His current research interest include: the modeling of fault risk, machine learning, social networks, derivatives-pricing models, portfolio and venture capital. He has published over 70 articles in academic journals and his recent book *Derivatives: Principles and Practice* was published in May 2010.

Christian Davies is a senior quantitative research analyst at Macquarie Securities and specializes in style research, multifactor modeling and developing stock selection strategies. He previously worked on the Quant Team at Citi and prior to that Christian spent 8 years at Schroder Investment Management as an equity quant analyst as well as an analyst within the Asia Team.

Dan diBartolomeo is founder and president of Northfield Information Services, a provider of analytical models for the global institutional investment community. He is also a Visiting Professor at the CARISMA research centre of Brunel University. Dan has published a long list of books, book chapters, and papers in professional and academic journals.

Huu Nhan Duong is a senior lecturer in finance at the School of Accounting, Economics and Finance, Deakin University, Australia. Dr. Duong's research interests are in the areas of market microstructure, derivatives market, and corporate finance. He has

published in the *Journal of Banking and Finance*, the *Journal of Futures Markets*, and the *Pacific-Basin Finance Journal.*

Michal Dzielinski is currently working towards his PhD at the Swiss Banking Institute under the supervision of Prof. Thorsten Hens. His research focus is on quantifying the impact of incoming news stories on the stock market for applications in financial modelling. His research is part of an interdisciplinary project, involving researchers from finance, communication science, computer linguistics, as well as industry partners.

Armando Gonzalez is the co-founder and CEO of RavenPack and has established it as a premier firm in sentiment analysis and natural language processing. Armando is widely regarded as one of the most knowledgeable authorities on automated news and sentiment analysis. His commentary has appeared in leading business media such as the *Wall Street Journal*, Dow Jones Newswires, CNBC, *The Trade News*, among others. Armando is a recognized speaker at conferences on behavioral finance and algorithmic trading across the globe.

Peter Hafez is the Director of Quantitative Research, RavenPack. A graduate and researcher from Sir John Cass Business School, Peter has held various positions in the portfolio management and alternative investment industry with companies such as Standard & Poor's, Credit Suisse First Boston, and Saxo Bank where he was Chief Quantitative Analyst and Head of CHARM. In 2008 he joined RavenPack as Director of Quantitative Research.

Alexander D. Healy conducts applied research at AlphaSimplex Group, with a focus on risk management, asset allocation, and nonparametric investment models. Alex holds an AB in mathematics and computer science (2002) and a PhD in theoretical computer science (2007), both from Harvard University. His dissertation research focused on the uses of randomness and randomized processes in algorithms and cryptography and, in particular, introduced new methods for generating pseudorandom numbers along with new applications of these methods.

Petko Kalev is an associate professor and Head of Research at the School of Commerce, University of South Australia. Dr. Kalev is an expert in empirical and applied finance and, specifically, in market microstructure, with a background in mathematics, statistics, and econometrics. His current research interests comprise capital markets/market microstructure, corporate finance, corporate governance, market efficiency, investments/funds management, and behavioral finance.

John W. Kittrell is a quantitative analyst at Knightsbridge Asset Management in Newport Beach, CA. John was a recipient of the National Science Foundation VIGRE Fellowship while at UCLA and was a guest lecturer at the joint CalTech–UCLA Logic Seminars in 2006 and 2007. His academic work has appeared in such publications as the *Proceedings of the American Mathematical Society* and *Ergodic Theory and Dynamical Systems.*

David Leinweber is the author of *Nerds on Wall Street: Maths, Machines and Wired Markets* (Wiley, 2009). He is Director of the Center for Innovative Financial Technology at Berkeley National Lab in Berkeley, CA. His professional interests focus on how modern information technologies are best applied in trading and investing and how

technology affects global financial markets. As a founder of two financial technology companies, and as a quantitative investment manager, he is an active participant in today's transformation of markets.

Andrew W. Lo is the Harris & Harris Group Professor of Finance at the MIT Sloan School of Management, Director of MIT's Laboratory for Financial Engineering, and founder and Chief Investment Strategist of the investment advisory firm AlphaSimplex Group, LLC. He has published numerous articles in finance and economics journals and has authored several books including *The Econometrics of Financial Markets*, *A Non-Random Walk Down Wall Street*, and *Hedge Funds: An Analytic Perspective*.

Andy Moniz, CFA is a senior quantitative research analyst at Macquarie Securities. His interests include statistical pattern recognition, Bayesian classifiers, event-driven strategies, stock selection, and style-timing research. He previously worked at Citi as part of the #1-ranked European Quantitative Research Team. Andy began his career on the Forecast Team at the Bank of England and also worked as a strategist within fundamental research at Credit Suisse.

Marian Munz is the founder, President, and Chief Executive Officer of Media Sentiment, Inc. He invented the media sentiment concept and technology and led development of the proprietary technology that delivers consistent results. Munz is one of the world's experts on financial news and media analysis, internet software, and decision support systems

Terrance Odean is the Rudd Family Foundation Professor of Finance at the Haas School of Business at the University of California, Berkeley. He is Chair of the Haas School's finance group, an associate editor at the *Journal of Finance*, a member of the *Journal of Investment Consulting* editorial advisory board, of the Russell Sage Behavioral Economics Roundtable, and of the Russell Investments Academic Advisory Board. His current research focus is on how psychologically motivated decisions affect investor welfare and security prices.

Marc Oliver Rieger is a full professor at the University of Trier. His recent research focuses on behavioral finance, especially investor behavior. He is author of two books: one on derivatives and one on financial economics.

Jacob Sisk is a principal at Leinweber & Co. and founder of Infoshock Inc. A former senior research scientist at Yahoo! he has been active in applying textual analytics, machine learning, and social network analysis to investment and trading for over 10 years. Jacob attended Reed College and holds advanced degrees in math and business from Tufts University and UCLA.

Adam Strudwick is a senior quantitative research analyst at Macquarie Securities and focuses his research on portfolio construction, implementation issues and multifactor modeling. He previously worked on the Quant Team at Citi; he also worked as an equity quant analyst at ABN Amro and before that as a management consultant with Accenture.

Tõnn Talpsepp, PhD, CFA holds a senior researcher position at Tallinn University of Technology and is involved in behavioral finance, volatility, and financial markets–related research in collaboration with working groups at the University of Trier and

the University of Zurich. Dr. Talpsepp was previously employed in the Risk Management Department of Swedbank and is currently involved in the trading and research activities of a proprietary trading firm.

Mark P.W. Vreijling is the R&D Director and co-founder of SemLab and has more than 15 years' experience in research and high-technology product development. Dr. Vreijling is a scientist at heart with a clear focus on the practical opportunities of scientific innovation.

Abbreviations and acronyms

ADR	American Depository Receipt
AI	Artificial Intelligence
AMEX	American Stock Exchange (now NYSE Amex Equities)
AMH	Adaptive Market Hypothesis
APARCH	Asymmetric Power GARCH (Generalized AutoRegressive Conditional Heteroskedasticity)
API	Application Programming Interface
APT	Arbitrage Pricing Theory
ARCH	AutoRegressive Conditional Heteroskedasticity
ARMA	AutoRegressive Moving Average
ASE	American Stock Exchange
ASX	Australian Stock eXchange
CAPM	Capital Asset Pricing Model
CAR	Cumulative Abnormal Return
CPD	Cumulative Probability Distribution
CRS	Company Relevance Score
CRSP	Center for Research in Security Prices
CSV	Comma Separated Value
CUVOALD	Computer Usable Version of the *Oxford Advanced Learner's Dictionary*
CVaR	Conditional Value at Risk
DJNA	Dow Jones News Analytics
EDA	Exploratory Data Analysis
EDGAR	Electronic Data Gathering, Analysis and Retrieval
EGARCH	Exponential Generalized AutoRegressive Conditional Heteroskedasticity
EMH	Efficient Market Hypothesis
ENS	Event Novelty Score
EPS	Earnings Per Share
ES	Expected Shortfall
ESS	Event Sentiment Score
ETF	Exchange Traded Fund
FD	Full Disclosure

FXR	Foreign eXchange Related
FX	Foreign eXchange
GARCH	Generalized AutoRegressive Conditional Heteroskedasticity
GI	General Inquirer
IA	Intelligence Amplification
ICB	Industry Classification Benchmark
IFRS	International Financial Reporting Standard
IG	Information Gain
IPO	Initial Public Offering
ISIN	International Securities Identification Number
MDH	Mixture of Distribution Hypothesis
ML	Machine Learning
MSCI	Morgan Stanley Capital International
MSH	Morgan Stanley High Tech Index
NA	News Analytics
NASDAQ	National Association of Securities Dealers Automated Quotations
NAV	Net Asset Value
NC	Naive Classifier
NEI	NewsScope Event Index
NLP	Natural Language Processing
NORM	News Optimized Risk Management
NVWAP	News Volume Weighted Average Price
NYSE	New York Stock Exchange
OLS	Ordinary Least Squares
PL/I	Programming Language I
PPI	Producers' Price Index
RIC	Reuters Instrument Code
RNSE	Reuters NewsScope Sentiment Engine
RSS	Rich Site Summary
SEC	Securities and Exchange Commission
SIRCA	Securities Industry Research Centre for Australasia
SVM	Support Vector Machine
SWAG	"Scientific Wild Ass Guess"
TRBC	Thomson Reuters Business Classification
TRNA	Thomson Reuters News Analytics
UTC	Coordinated Universal Time
VaR	Value at Risk
VC	Venture Capital
VWAP	Volume Weighted Average Price
XBRL	Extensible Business Reporting Language
XML	Extensible Markup Language

Contents