Text Data in Finance

Professor: Harry Mamaysky
TA: Jack McCoy

Spring 2024

1 Overview

The last several years have seen an enormous increase in the use of text data to supplement traditional financial decision making. For example, all major sell side and buy side firms, as well as financial data providers, like S&P Global, Moody's, and Bloomberg, now devote significant resources to the processing of text data sets. These include news, earnings calls, Fed and other central bank communications, and social media. The major technology firms have similarly made large investments in natural language processing (NLP), which has produced many open-source packages (such as TensorFlow from Google and PyTorch from Facebook) that have attracted large user bases in industry and academia. There is, in addition, a vibrant ecosystem of start-up firms focused on innovative uses of text data in finance, such as RavenPack, Amenity Analytics, and Aiera, to name just a few.

While NLP methods are now well developed, their application to finance and economics is more nascent (see Gentzkow, Kelly, and Taddy 2019). This course introduces students to state of the art NLP methods and their applications to traditional finance problems. The course is Python-based, analytically rigorous, and emphasizes the use of open-source NLP libraries. While there are other NLP courses at Columbia, this course is unique because of its focus on the application of NLP methods to problems in finance and economics.

PhD students will find the material valuable for their research. MS students will find the course valuable because firms, including many hedge funds, are looking for employees who understand finance and can work with text data. Quantitatively oriented MBA students will find the course valuable because it will prepare them to lead teams and

build products that use text data.

2 Course structure

We cover the basics of text processing, including stemming, tokenization, number conversions, dropping stop words, and constructing document-term matrixes. We then do basic sentiment calculations using existing sentiment dictionaries, and also discuss recent techniques that allow word tone to be determined from combined market and text data. Topic modeling is then introduced as a way of extracting deeper structure from text. We next cover neural network (NN) and word embedding methods, and apply these to calculate the unusualness of text. The course finishes with a discussion of economic narratives as a unifying theme for how markets respond to text information. Importantly, all NLP tools are analyzed in the context of solving financial problems.

Our two main data sets are the Reuters news archive, a collection of over 70 million news articles carried by Reuters since 1996, and the S&P Global Transcripts database, a collection of hundreds of thousands of corporate earnings calls dating back to 2008.

The course has weekly, coding-intensive problem sets focused on using NLP to analyze text data to better understand financial markets. The problem sets are motivated by recent published or current working papers in finance and economics. The course emphasizes rigorous econometric analysis of financial data sets using traditional information, as well as information that can be gleaned only from text data.

Students are expected to have proficiency in Python (though course work can also be done in R). Most of the text data that will be analyzed will be presented in already cleaned form.

3 Schedule

This is a half-term course which will consist of the following sessions. Each session represents a three hour class.

Class	Date	Topic	HW	Due Date
1	3/22	Introduction & early attempts	1	4/05
2	3/29	Sentiment, underreaction, and trading costs		
3	4/05	Learning from the data	2	4/19
4	4/12	Topic modeling		
5	4/19	Textual similarity	3	4/29 (Mon)
6	4/26	Neural networks: An introduction		

Homeworks

1. HW1: Merge data; sentiment and future returns; trading strategy
2. HW2: Learning from the data: implied word tone and topic models
3. HW3: Earnings calls similarity and neural network intro

4 References

The following is an incomplete list of relevant papers.

4.1 Early papers and surveys

*Antweiler, W., and Frank, M, 2004, "Is all that talk just noise? The information content of internet stock message boards," *Journal of Finance*, 59 (3), 1259–1294.

*Das, S. and M, Chen, 2007, "Yahoo! for Amazon: Sentiment extraction from small talk on the web," *Management Science*, 53 (9), 1375–1388.

Gentzkow, M., B. Kelly, and M. Taddy, 2019, "Text as data," *Journal of Economic Literature*, 57 (3), 535–574.

4.2 Sentiment

*Garcia, D., 2013, "Sentiment during recessions," *Journal of Finance*, 68 (3), 1267–1300.

Gross-Klussman, A. and N. Hautsch, 2011, "When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions," *Journal of Empirical Finance*, 18, 321–340.

*Loughran, T. and B. McDonald, 2011, "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks," *Journal of Finance*, 66, 35–65.

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

4.3 Underreaction

Frank, M. and A. Sanati, 2018, "How does the stock market absorb shocks," *Journal of Financial Economics*, 129, 136–153.

Glasserman, P., F. Li, and H. Mamaysky, 2022, "Time Variation in the News-Returns Relationship," R&R *JFQA*.

*Heston, S. and N. Sinha, 2017, "News vs. sentiment: Predicting stock returns from news stories," *Financial Analysts Journal*, 73 (3), pp. 67–83.

*Ke, Z.T., B. Kelly, and D. Xiu, 2019, "Predicting returns with text data," working paper.

Jiang, J., B. Kelly, and D. Xiu, 2023, "Expected returns and large language models," working paper.

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

Theories of Underreaction

Daniel, K., D. Hirshleifer, and A. Subrahmanyam, 1998, "Investor Psychology and Security Market Under- and Overreactions," *Journal of Finance*, 53 (6), 1839–1885.

Hong, H., and J. Stein, 1999, "A unified theory of underreaction, momentum trading, and overreaction in asset markets," *The Journal of Finance*, 54 (6), 2143–2184.

4.4 Trading costs

*add trading cost papers...

4.5 Learning from data

*Garcia, D., X. Hu, and M. Rohrer, 2020, "The color of finance words," working paper.

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

Kelly, B., A. Manela, and A. Moreira, 2021, "Text selection," *Journal of Business and Economic Statistics*, 39 (4), 859–879.

*Manela, A. and A. Moreira, 2017, "News implied volatility and disaster concerns," *Journal of Financial Economics*, 123 (1), 137–162.

*Taddy, M., 2013, "Multinomial inverse regression for text analysis," *Journal of American Statistical Association*, 755–770.

4.6 Topic analysis

Asuncion, A., M. Welling, P. Smyth, and Y. Teh, 2009, "On smoothing and inference for topic models," *UAI '09: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 27–34.

> Comparison of variational inference vs. Gibbs sampling for estimating topic models.

Blei, D., A. Ng, and M. Jordan, 2003, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, 3, 993–1022.

Original LDA paper with variational inference.

Glasserman, P., K. Krstovski, P. Laliberte, and H. Mamaysky, 2020, "Choosing news topics to explain stock market returns," *Proceedings of ACM International Conference on AI in Finance (ICAIF '20)*.

*Griffiths, T. and M. Steyvers, 2004, "Finding scientific topics," *Proceedings of the National Academy of Sciences* , 101 (suppl 1), 5228–5235.

Collapsed Gibbs sampler.

Ke, S., J.L. Montiel Olea, and J. Nesbit, 2020, "A robust machine learning algorithm for text analysis," working paper.

Indeterminacy of LDA.

4.7 Topic analysis – applications

*Bybee, L., B. Kelly, A. Manela, and D. Xiu, 2021, "Business news and business cycles," working paper.

Bybee, L., B. Kelly, Y. Su, 2022, "Narrative asset pricing: Intepretable systemic risk factors from news text," working paper.

Calomiris, C. and H. Mamaysky, 2019, "How news and its context drive risk and returns around the world," *Journal of Financial Economics*, 133 (2), 299–336.

4.8 Textual similarity

*Cohen, L., C. Malloy, and Q. Nguyen, 2020, "Lazy prices," *Journal of Finance*, 75 (3), 1371–1414.

*Hoberg and Philips, 2016, "Text-based network industries and endogenous product differentiation," *Journal of Political Economy*.

*Tetlock. P., 2011, "All the news that's fit to print: Do markets overreact to stale information?" *Review of Financial Studies*, 24 (5), 1481–1512.

4.9 Neural networks

*Charniak, E., 2018, *Introduction to Deep Learning*, MIT Press.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, 2017, "Attention is all you need," *31st Conference on Neural Information Processing Systems (NIPS 2017)*.

> Introduces the *Transformer*, an architecture similar to recurrent neural networks, but more efficient and easier to parallelize. There is an annotatd version of the paper (including code) here.

Radford, A., K. Narasimhan, T. Salimans, and I. Sutskever, 2018, "Improvinig language understanding by generative pre-training," *Working Paper OpenAI*.

> Introduces the GPT model, which is based on the transfomer architecture.

Devlin, J., M.-w. Chang, K. Lee, and K. Toutanova, 2019, "BERT: Pre-training of deep bidirectional transformers ," *Working Paper Google AI*.

> Introduces BERT model, which is also based on the Transformer architecture.

4.10 Economic policy uncertainty and entropy

Baker, S., N. Bloom, and S. Davis, 2016, "Measuring economic policy uncertainty," *Quarterly Journal of Economics*, 131 (4), 1593–1636.

Brogaard, J. and A. Detzel, 2015, "The asset-pricing implications of economic policy uncertainty," *Management Science*, 61 (1), 3–18.

Glasserman, P. and H. Mamaysky, 2019, "Does unusual news forecast market stress?" *Journal of Financial and Quantitative Analysis*, 54 (5), 1937–1974.

Glasserman, P., H. Mamaysky, and J. Qin, 2022, "The price of entropy risk," working paper.

4.11 Narratives

Garcia, D., 2018, "The kinks of financial journalism," working paper.

Goetzmann, W., D. Kim, and R. Shiller, 2022, "Crash narratives," working paper.

Mamaysky, H., 2022, "News and markets in the time of COVID-19," R&R *JFQA*.

Cookson, J.A., J. Engelberg, and W. Mullins, 2021, "Echo chambers," working paper.

Shiller, R., 2017, "Narrative economics," *American Economic Review*, 107 (4), 967–1004.