Text Mining in Economics

Stephen Hansen, stephen.hansen@imperial.ac.uk

1 Textbooks / Overview Material

There is no one source that covers all of the material in the course. Grimmer and Stewart (2013), Bholat et al. (2015), and Gentzkow et al. (2019a) are survey articles that provide accessible introductions to text mining.

2 Lecture 1: Introduction to Text Data

- Tetlock (2007), Loughran and Mcdonald (2011), Shapiro et al. (2020), Nyman et al. (2021)
- Baker et al. (2016)
- Hassan et al. (2019)

3 Lecture 2: Word Embeddings

- Deerwester et al. (1990)
- Mikolov et al. (2013a,b)
- Ash et al. (2020)
- Hansen et al. (2021)

4 Lecture 3: Text Regression

- Taddy (2013, 2015)
- Gentzkow et al. (2019b)
- Davis et al. (2020)

5 Lecture 4: Latent Variable Models

- Blei et al. (2003)
- Hansen et al. (2018)

- Mueller and Rauh (2018)
- Larsen and Thorsrud (2019), Thorsrud (2020)
- Bandiera et al. (2020)
- Roberts et al. (2014, 2016)

References

Ash, E., Chen, D. L., and Ornaghi, A. (2020). Gender attitudes in the judiciary : Evidence from U.S. circuit courts. https://warwick.ac.uk/fac/soc/economics/research/workingpapers/2020/twerp_1256_-_ornaghi.pdf.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636.

Bandiera, O., Prat, A., Hansen, S., and Sadun, R. (2020). CEO Behavior and Firm Performance. *Journal of Political Economy*, 128(4):1325–1369.

Bholat, D., Hans, S., Santos, P., and Schonhardt-Bailey, C. (2015). *Text Mining for Central Banks*. Centre for Central Banking Studies, Bank of England.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3(null):993–1022.

Davis, S. J., Hansen, S., and Seminario-Amez, C. (2020). Firm-Level Risk Exposures and Stock Returns in the Wake of COVID-19. Working Paper 27867, National Bureau of Economic Research.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

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

Gentzkow, M., Shapiro, J. M., and Taddy, M. (2019b). Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica*, 87(4):1307–1340.

Grimmer, J. and Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297.

Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach. *The Quarterly Journal of Economics*, 133(2):801–870.

Hansen, S., Ramdas, T., Sadun, R., and Fuller, J. (2021). The Demand for Executive Skills. Technical Report 28959, National Bureau of Economic Research, Inc.

Hassan, T. A., Hollander, S., van Lent, L., and Tahoun, A. (2019). Firm-Level Political Risk: Measurement and Effects. *The Quarterly Journal of Economics*, 134(4):2135–2202.

Larsen, V. H. and Thorsrud, L. A. (2019). The value of news for economic developments. *Journal of Econometrics*, 210(1):203–218.

Loughran, T. and Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. *arXiv:1310.4546 [cs, stat]*.

Mueller, H. and Rauh, C. (2018). Reading Between the Lines: Prediction of Political Violence Using Newspaper Text. *American Political Science Review*, 112(2):358–375.

Nyman, R., Kapadia, S., and Tuckett, D. (2021). News and narratives in financial systems: Exploiting big data for systemic risk assessment. *Journal of Economic Dynamics and Control*, 127:104119.

Roberts, M. E., Stewart, B. M., and Airoldi, E. M. (2016). A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association*, 111(515):988–1003.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, 58(4):1064–1082.

Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2020). Measuring news sentiment. *Journal of Econometrics*.

Taddy, M. (2013). Multinomial Inverse Regression for Text Analysis. *Journal of the American Statistical Association*, 108(503):755–770.

Taddy, M. (2015). Distributed Multinomial Regression. *The Annals of Applied Statistics*, 9(3):1394–1414.

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, 62(3):1139–1168.

Thorsrud, L. A. (2020). Words are the New Numbers: A Newsy Coincident Index of the Business Cycle. *Journal of Business & Economic Statistics*, 38(2):393–409.