Text analytics, NLP, and accounting research

2020 November 11

Dr. Richard M. Crowley
rcrowley@smu.edu.sg
http://rmc.link/

Foundations

What is text analytics?

> Extracting meaningful information from text

- This could be as simple as extracting specific words/phrases/sentences
- This could be as complex as extracting latent (hidden) patterns structures within text
  - Sentiment
  - Content
  - Emotion
  - Writer characteristics
  - …
- Often called text mining (in CS) or textual analysis (in accounting)

What is NLP then?

> NLP is a field devoted to understanding how to understand human language

- NLP stands for **N**atural **L**anguage **P**rocessing
- It is a very diverse field within CS
  - Grammar/linguistics
  - Conversations
  - Conversion from audio, images
  - Translation
  - Dictation
  - Generation

Why discuss NLP?

Consider the following situation:

> You have a collection of 1 million sentences, and you want to know which are accounting relevant

- Without NLP:
  1. Hire an RA/mechanical turk army…
  2. Use a dictionary: Words/phrases like “earnings,” “profitability,” “net income” are likely to be in the sentences
- With NLP:
  1. We could associate sentences with outside data to build a classifier (supervised approach)
  2. We could ask an algorithm to learn the structure of all sentences, and then extract the useful part *ex post* (unsupervised)

Data that has been studied

- Firms
  - Letters to shareholders
  - Annual and quarterly reports
  - 8-Ks
  - Press releases
  - Conference calls
  - Firm websites
  - Twitter posts
- Investors
  - Blog posts
  - Social media posts
- Intermediaries
  - Newspaper articles
  - Analyst reports
- Government
  - FASB exposure drafts
  - Comment letters
  - IRS code
  - Court cases

A brief history of text analytics in accounting research

Cole and Jones 2005 Jones and Shoemaker 1994

1980s and 1990s

Manual content analysis

- Read through “small” amounts of text, record selected aspects

Indexes

- Ex.: Botosan (1997 TAR): For firms with low analyst following, more disclosure ⇒ Lower cost of equity
  - Index of 35 aspects of 10-Ks
- Covered in detail in Cole and Jones (2004 JAL)
  - Most use small samples
  - Often use select industries

Readability

- Automated starting with Dorrell and Darsey (1991 JTWC) in accounting…
- At least 32 studies on this in the 1980s and early 1990s per Jones and Shoemaker (1994 JAL)
  - Only 2 use full docs
  - Only 2 use >100 docs

2000s

Automation

- With computer power increasing, two new avenues opened:
  1. Do the same methods as before, at scale
     - Ex.: Li (2008 JAE): Readability, but with many documents instead of <100
  2. Implementing statistical techniques (often for tone/sentiment)
     - For instance, sentiment classification with Naïve Bayes, SVM, or other statistical classifiers
       - Antweiler and Frank (2005 JF)
       - Das and Chen (2007 MS)
       - Li (2010 JAR)

Early 2010s

> Dictionaries take the helm

- Loughran and McDonald (2011 JF) points out the misspecification of using dictionaries from other contexts
  - Also provides a sets of positive, negative, modal strong/weak, litigious, and constraining words (available here)
- Subsequent work by the authors provides a critique:

> Applying financial dictionaries "without modification to other media such as earnings calls and social media is likely to be problematic" (Loughran and McDonald 2016)

- A lot of papers ignore this critique, and are still at risk of *misspecification*

Late 2010s to present

Fragmentation and new methods

- Loughran and McDonald dictionaries frequently used
- Bog index is perhaps a new entrant in the Fog index vs document length debate
- LDA methods first published in Accounting/Finance in Bao and Datta (2014 MS), with a handful of other papers following suit.
- More methods on the horizon

Going forward

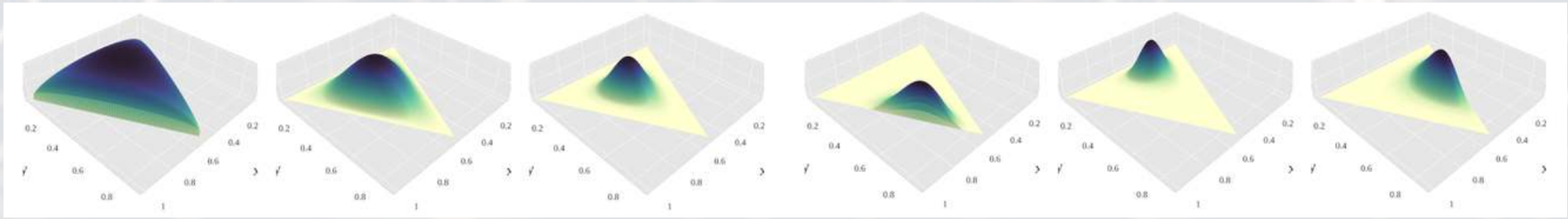
> A lot of choices

- Why? Because accounting research has been behind the times, but seems to be catching up
  - We can incorporate more than a year's worth of innovation in NLP each year…

Useful methods for analytics

Content classification: Latent Dirichlet Allocation

- *L*atent *D*irichlet *A*llocation, from Blei, Ng, and Jordan (2003)
- One of the most popular methods under the field of *topic modeling*
- LDA is a Bayesian method of assessing the content of a document
- LDA assumes there are a set of topics in each document, and that this set follows a *Dirichlet* prior for each document
- Words within topics also have a *Dirichlet* prior



More details from the creator

Example: LDA, 10 topics, all 2014 10-Ks

```
# Topics generated using R's stm library
labelTopics(topics)
```

```
## Topic 1 Top Words:
##       Highest Prob: properti, oper, million, decemb, compani, interest, leas
##       FREX: ffo, efih, efh, tenant, hotel, casino, guc
##       Lift: aliansc, baluma, change-of-ownership, crj700s, directly-reimburs, e
##       Score: reit, hotel, game, ffo, tenant, casino, efih
## Topic 2 Top Words:
##       Highest Prob: compani, stock, share, common, financi, director, offic
##       FREX: prc, asher, shaanxi, wfoe, eit, hubei, yew
##       Lift: aagc, abramowitz, accello, akash, alix, alkam, almati
##       Score: prc, compani, penni, stock, share, rmb, director
## Topic 3 Top Words:
##       Highest Prob: product, develop, compani, clinic, market, includ, approv
##       FREX: dose, preclin, nda, vaccin, oncolog, anda, fdas
##       Lift: 1064nm, 12-001hr, 25-gaug, 2ml, 3shape, 503b, 600mg
##       Score: clinic, fda, preclin, dose, patent, nda, product
## Topic 4 Top Words:
##       Highest Prob: invest, fund, manag, market, asset, trade, interest
##       FREX: uscf, nfa, unl, uga, mlai, bno, dno
##       Lift: a-1t, aion, apx-endex, bessey, bolduc, broyhil, buran
##       Score: uscf, fhlbank, rmbs, uga, invest, mlai, ung
## Topic 5 Top Words:
```

Papers using LDA (or variants)

- Bao and Datta (2014 MS): Quantifying risk disclosures
- Bird, Karolyi, and Ma (2018 working): 8-K categorization mismatches
- Brown, Crowley, and Elliott (2020 JAR):
  - Content based fraud detection
- Crowley (2018 working):
  - Mismatch between 10-K and website disclosures
- Crowley, Huang, and Lu (2020 working 1; 2020 working 2):
  - Financial and executive disclosure on Twitter
- Crowley, Huang, Lu, and Luo (2019 working):
  - CSR disclosure on Twitter
- Dyer, Lang, and Stice-Lawrence (2017 JAE):
  - Changes in 10-Ks over time
- Hoberg and Lewis (2017 JCF): AAERs and 10-K MD&A content, ex post
- Huang, Lehavy, Zang, and Zheng (2018 MS):
  - Analyst interpretation of conference calls

Sentiment: Varied

- General purpose word lists like Harvard IV
  - Tetlock (2007 JF)
  - Tetlock, Saar-Tsechansky, and Macskassy (2008 JF)
- Many recent papers use *10-K specific* dictionaries from Loughran and McDonald (2011 JF)
- Some work using Naive Bayes and similar
  - Antweiler and Frank (2005 JF), Das and Chen (2007 MS), Li (2010 JAR), Huang, Zang and Zheng (2014 TAR), Sprenger, Tumasjan, Sandner, and Welpe (2014 EFM)
- Some work using SVM
  - Antweiler and Frank (2005 JF)

Sentiment: What is used in practice (CS side)

> "The prevalence of polysemes in English – words that have multiple meanings – makes an absolute mapping of specific words into financial sentiment impossible." – Loughran and McDonald (2011)

- Embedding methods *can* make this possible
- Embeddings abstract away from words, converting words/ phrases/ sentences/ paragraphs/ documents to high dimensional vectors
  - Used in Brown, Crowley, and Elliott (2020 JAR) (word level)
  - Used in Crowley, Huang, and Lu (2020 Working 2) (sentence/document level)
- Embeddings are passed to a supervised classifier to learn sentiment
- Other methods include weak supervision
  - Such as the *Joint Sentiment Topic* model by Lin and He (2009 ACM) (used in Crowley (2018 working))

Readability…

- 2008: Fog index kick-started this area in accounting
  - Li (2008 JAE), a bunch of other papers
- 2014: File length captures complexity more accurately…
  - Loughran and McDonald (2014 JF; 2016 JAR)
- 2017: Bog index
  - Bonsall, Leone, Miller and Rennekamp (2017 JAE); Bonsall and Miller (2017 RAST)
  - Subject to Loughran and McDonald's critique of general purpose dictionaries

> "[…] The use of word lists derived outside the context of business applications has the potential for errors that are not simply noise and can serve as unintended measures of industry, firm, or time period. The computational linguistics literature has long emphasized the importance of developing categorization procedures in the context of the problem being studied (e.g., Berelson [1952])." – LM 2016

Readability…

> “There are problems with the face validity of the accounting readability studies. Accounting researchers have, in general, assumed that the readability formulas measure not only readability but also understandability. Indeed, readability and understandability have often been used interchangeably, the assumption being they are synonymous. However, although these concepts are related, they do differ.” – Jones and Shoemaker (1994 JAL)

The literature has not *yet* addressed this.

Going forward

Going forward

- There are a lot of cool methods
  - There are a lot of cool measures
- It is easy to get wrapped up in the technical details and achievements and lose sight of the *purpose* for using them

> Tailor-made measures

- Tone dispersion (Allee and DeAngelis 2015 JAR)
- Disclosure “Scriptability” (Allee, DeAngelis, and Moon 2018 JAR)
- Content differences
  - DeAngelis (2014 dissertation) – unique content
  - Crowley (2018 working) – extent of content differences
- Industry classification
  - Hoberg and Phillips (6 papers)

Recommended coding libraries

Python:

- Text parsing: `spaCy`
- LDA: `gensim`
- Sentiment: `NLTK`, `SpaCy`, or handcode using `Counter()` (super fast)
- Classifiers: `scikit-learn` or `keras` or `pytorch` or `huggingface`
- Other measures: `NLTK`, `spaCy`

R:

- LDA: `stm` + `quanteda` + `convert(dfm,to='stm')`
- Sentiment (dictionary): `tidytext`
- Classifiers: `caret`, `e1071`, or `keras`
- Other measures: Using python is likely better

- Also useful: MALLET, Stanford NLP

References

- Allee, Kristian D., and Matthew D. DeAngelis. 2015. “The Structure of Voluntary Disclosure Narratives: Evidence from Tone Dispersion.” Journal of Accounting Research 53 (2): 241–74. https://doi.org/10.1111/1475-679X.12072.
- Allee, Kristian D., Matthew D. DeAngelis, and James R. Moon. 2018. “Disclosure ‘Scriptability.’” Journal of Accounting Research 56 (2): 363–430. https://doi.org/10.1111/1475-679X.12203.
- Antweiler, Werner, and Murray Z. Frank. 2005. “Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards.” The Journal of Finance 59 (3): 1259–94. https://doi.org/10.1111/j.1540-6261.2004.00662.x.
- Y. Bao, and A. Datta. 2014. “Simultaneously Discovering and Quantifying Risk Types from Textual Disclosures.” Management Science 60 (6): 1371–1391.
- Bird, Andrew, Stephen A. Karolyi, and Paul Ma. 2018. “Strategic Disclosure Misclassification.” SSRN Scholarly Paper ID 2778805. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2778805.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. “Latent Dirichlet Allocation.” J. Mach. Learn. Res. 3 (March): 993–1022.
- Bonsall, Samuel B., Andrew J. Leone, Brian P. Miller, and Kristina Rennekamp. 2017. “A Plain English Measure of Financial Reporting Readability.” Journal of Accounting and Economics 63 (2): 329–57. https://doi.org/10.1016/j.jacceco.2017.03.002.
- Bonsall, Samuel B., and Brian P. Miller. 2017. “The Impact of Narrative Disclosure Readability on Bond Ratings and the Cost of Debt.” Review of Accounting Studies 22 (2): 608–43. http://dx.doi.org.libproxy.smu.edu.sg/10.1007/s11142-017-9388-0.
- Botosan, C. A. 1997. “Disclosure level and the cost of equity capital.” The Accounting Review 72 (3), 323–349.
- Brown, Nerissa C., Richard Crowley, and W. Brooke Elliott. 2020. “What Are You Saying? Using Topic to Detect Financial Misreporting.” Journal of Accounting Research.
- Cole, C. J. and C. L. Jones. “Management Discussion and Analysis: A Review and Implications for Future Research.” Journal of Accounting Literature 24, 135–174.

References

- Crowley, Richard. 2018. “Disclosure through Multiple Disclosure Channels.”
- Crowley, Richard, Wenli Huang, and Hai Lu. 2020 (1). “Discretionary Disclosure on Twitter.” SSRN Scholarly Paper ID 3105847. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3105847.
- Crowley, Richard, Wenli Huang, and Hai Lu. 2020 (2). “Executive Tweets.” Working Paper
- Crowley, Richard, Wenli Huang, Hai Lu, and Wei Luo. 2019. “Do Firms Manage Their CSR Reputation? Evidence from Twitter.” Working paper, Singapore Management University.
- Das, Sanjiv R., and Mike Y. Chen. 2007. “Yahoo! For Amazon: Sentiment Extraction from Small Talk on the Web.” Management Science 53 (9): 1375–88. https://doi.org/10.1287/mnsc.1070.0704.
- Dorrell, J. T., and N. S. Darsey. 1991. “An analysis of the readability and style of letters to stockholders.” Journal of Technical Writing and Communication 21: 73–83.
- Dyer, Travis, Mark Lang, and Lorien Stice-Lawrence. 2017. “The Evolution of 10-K Textual Disclosure: Evidence from Latent Dirichlet Allocation.” Journal of Accounting and Economics 64 (2): 221–45. https://doi.org/10.1016/j.jacceco.2017.07.002.
- Hoberg, Gerard, and Craig Lewis. 2017. “Do Fraudulent Firms Produce Abnormal Disclosure?” Journal of Corporate Finance 43 (April): 58–85. https://doi.org/10.1016/j.jcorpfin.2016.12.007.
- Huang, Allen H., Reuven Lehavy, Amy Y. Zang, and Rong Zheng. 2018. “Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach.” Management Science 64 (6): 2833–55. https://doi.org/10.1287/mnsc.2017.2751.
- Huang, Allen H., Amy Y. Zang, and Rong Zheng. 2014. “Evidence on the Information Content of Text in Analyst Reports.” Accounting Review 89 (6): 2151–80. https://doi.org/10.2308/accr-50833.
- Jones, M. J. and P. A. Shoemaker. “Accounting Narratives: A Review of Empirical Studies of Content and Readability.” Journal of Accounting Literature 13, 142.

References

- Li, Feng. 2008. “Annual Report Readability, Current Earnings, and Earnings Persistence.” Journal of Accounting and Economics, Economic Consequences of Alternative Accounting Standards and Regulation, 45 (2): 221–47. https://doi.org/10.1016/j.jacceco.2008.02.003.
- Li, Feng. 2010a. “The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach.” Journal of Accounting Research 48 (5): 1049–1102. https://doi.org/10.1111/j.1475-679X.2010.00382.x.
- Li, Feng. 2010b.
- Lin, Chenghua, and Yulan He. 2009. “Joint Sentiment/Topic Model for Sentiment Analysis.” In Proceedings of the 18th ACM Conference on Information and Knowledge Management, 375–384. CIKM ’09. New York, NY, USA: ACM. https://doi.org/10.1145/1645953.1646003.
- Loughran, Tim, and Bill McDonald. 2011. “When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks.” The Journal of Finance 66 (1): 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x.
- Loughran, Tim, and Bill McDonald. 2014. “Measuring Readability in Financial Disclosures.” The Journal of Finance 69 (4): 1643–71. https://doi.org/10.1111/jofi.12162.
- Loughran, Tim, and Bill McDonald. 2016. “Textual Analysis in Accounting and Finance: A Survey.” Journal of Accounting Research 54 (4): 1187–1230. https://doi.org/10.1111/1475-679X.12123.
- Sprenger, Timm O., Andranik Tumasjan, Philipp G. Sandner, and Isabell M. Welpe. 2014. “Tweets and Trades: The Information Content of Stock Microblogs.” European Financial Management 20 (5): 926–57. https://doi.org/10.1111/j.1468-036X.2013.12007.x.
- Tetlock, Paul C. 2007. “Giving Content to Investor Sentiment: The Role of Media in the Stock Market.” The Journal of Finance 62 (3): 1139–68. https://doi.org/10.1111/j.1540-6261.2007.01232.x.
- Tetlock, Paul C., Maytal Saar-Tsechansky, and Sofus Macskassy. 2008. “More Than Words: Quantifying Language to Measure Firms’ Fundamentals.” The Journal of Finance 63 (3): 1437–67. https://doi.org/10.1111/j.1540-6261.2008.01362.x.

Packages used for these slides

- kableExtra
- knitr
- revealjs
- stm