Text analytics, NLP, and accounting research

2018 November 23

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 **Natural Language Processing**
- 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
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
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

- Latent Dirichlet Allocation, 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, escena, hhmk
### 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, accelo, akash, alix, alka, 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, fhlibank, rmbs, uga, invest, mlai, ung

## Topic 5 Top Words:
### Highest Prob: servic, report, file, program, provid, network, requir
### FREX: techostar, fcc, fccs, telesat, tilec, starz, tretransmiss
### Lift: 1100-n, 2-usb, 2011-c1, 2012-ccre4, 2013-c9, aastra, accreditor
### Score: entergi, fcc, techostar, wireless, broadcast, video, cabl

## Topic 6 Top Words:
### Highest Prob: loan, bank, compani, financi, decemb, million, interest
### FREX: nonaccru, oreo, tdrs, bancorp, fdic, charge-off, alll
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 (2018 working):
  - Content based fraud detection
- Crowley (2016 working):
  - Mismatch between 10-K and website disclosures
- Crowley, Huang, and Lu (2018 working):
  - Financial disclosure on Twitter
- Crowley, Huang, Lu, and Luo (2018 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
- Some work using SVM
  - Antweiler and Frank (2005 JF)
Embeddings 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 (2018 working) (word level)
  - Used in WIP by Crowley, Huang, and Lu (sentence/document level)
- Embeddings are past 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 (2016 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 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 (2016 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
- 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

References


References

- Li, Feng. 2010b.