Harm Schütt, LMU München schuett@bwl.lmu.de

Thorsten Sellhorn, LMU München sellhorn@bwl.lmu.de

Automatisierte Textanalyse in der Rechnungslegungsforschung: Erkenntnisstand und Methodenfragen

Automated textual analysis (ATA) in accounting research

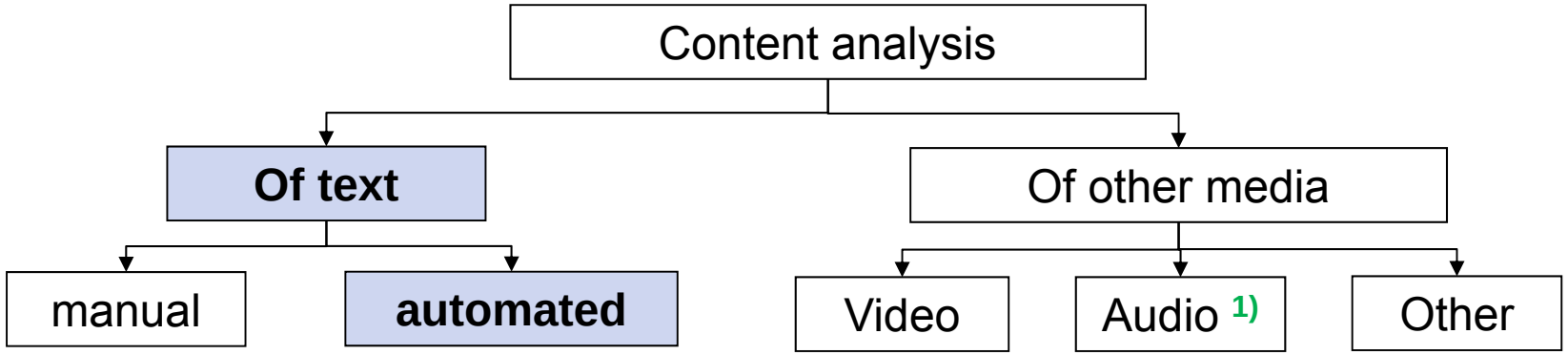
VHB WK RECH – Passau 2016
16. Februar 2016 | 13:15 – 14:45 Uhr

Agenda

1. Introduction	What is automated textual analysis (ATA hereafter), and why should we care?
2. Applications	What does ATA allow us to do?
3. Evidence	Which research questions are being addressed using ATA, and what has been found?
4. Methods	How does ATA work?
5. Critical discussion	What are the challenges of working with ATA?
6. Conclusion	What would we like you to take away from this talk?

What is automated textual analysis (ATA)?

- ATA extracts information by **parsing texts for patterns** using quantitative and automated methods rather than hand collection



- Synonyms and related terms: Quantitative content analysis, (statistical) natural language processing (NLP), information retrieval, computational linguistics, quantitative semantics, stylometrics, text mining

1) For example, Coval Shumway (2001 *JoF*); Hobson Mayew Venkatachalam (2012 *JoF*)

Why is ATA interesting for research in accounting?

ATA methods promise new insights into long-standing questions that occupy **disclosure and capital-markets** research:

- Does narrative disclosure (i.e., text) contain valuation-relevant information – beyond that inherent in the numbers?
  - Tetlock (2007 *JoF*): "linguistic .. content captures otherwise hard-to-quantify aspects of firms' fundamentals"
- Can information in narratives detect/predict important economic conditions/events, e.g., fraud or bankruptcy?
- Do firms "manage" text attributes strategically – like earnings?
- Does the way in which disclosures are written affect the way in which users process the signals being communicated?
  - IASB's concern about "information overload": *Disclosure Initiative*

Why is ATA research on the rise? [1)]

- ATA allows harvesting quantitative information from large bodies of text
- Facilitated by growing data processing, storage, transmission capacities
- Technical progress in computational linguistics and artificial intelligence (e.g., search engines, plagiarism software)
- Online availability of large text archives (e.g., SEC EDGAR, analysts' reports, conference call transcripts, press articles, social media)
- Growing focus on reproducibility of research – ATA less subjective?
- Increasingly applied in practice
  - SEC's Accounting Quality Model software ("RoboCop") [2)]
  - BlackRock fund „Global Long/Short Equity" uses algorithms to extract sentiment from analyst reports, earnings releases, and social media

1) See, for example, Loughran McDonald (2015 *SSRN*); Li (2008 *JoAccLit*) | **2)** Eaglesham (2013 *WSJ*)

Agenda

1. Introduction
2. Applications
3. Evidence
4. Methods
5. Critical discussion
6. Conclusion

ATA methods use …

A

… diverse sources of text …

B

… to generate different "outcomes"

1. **Text attributes**
   - "Sentiment" conveyed?
   - How readable?
   - Redundant or informative?
2. **Disclosure scores**
   - Overall quantity?
   - Content-specific quantity?
3. **Variables and datasets**
   - Risk factor disclosures?
   - Questions on fair value?

Text corpora generally lend themselves for ATA when they are:

- Highly machine-readable
  - HTML (+)
  - PDF (-)
- Large
- Correct orthography
  - Twitter, OCR‘ed documents (-)
  - Spoken language (-)

Other aspects specific to research question pursued

- Firm-originated vs external
- Spontaneous vs rehearsed/scripted
- Written vs spoken
- Voluntary vs mandatory disclosure
- Author known vs unknown
- Free vs costly
- Structured vs unstructured
- Standardized vs free-form
- Recurring vs ad-hoc
- Audited vs unaudited
- Regulated vs discretionary

A Text corpora used in extant research (selection)

Firm-originated disclosures

- 10-Ks or annual reports
  Loughran McDonald (2011 JoF) | Cazier Pfeiffer (2014 *SSRN*, 2015 *AccHor*) | You Zhang (2009 *RASt*) | Lang Stice-Lawrence (2015 *JAE*) | Chircop Tarsalewska (2015 *SSRN*) | Lundholm Li Minnis (2013 *JAR*)
- Specific 10-K sections
  MD&A: Brown Tucker (2011 *JAR*) | Humpherys Moffitt Burns Burgoon Felix (2011 *Decision Support Systems*) | | Davis Tama-Sweet (2012 *CAR*) | **Footnotes:** Peterson Schmardebeck Wilks (2015 *TAR*) | Chen Li (2013 *SSRN*) | **Product descriptions:** Hoberg Phillips (2010 *RFS*) | **Risk reporting:** Nelson Pritchard (2014 *SSRN*) | **CEO letter to shareholders:** Dikolli Keusch Mayew Steffen (2014 *SSRN*) | **Several:** Brown Knechel (2016 *SSRN*: Business description, footnotes, MD&A) | **Forward-looking statements:** Li (2010 *JAR*)
- 8-Ks or ad-hoc reports
  Cooper He Plumlee (2015 *SSRN*)
- Earnings announcements/press releases
  Henry (2008 *JBusComm*) | Rogers Van Buskirk Zechman (2011 *TAR*) | Davis Piger Sedor (2012 *CAR*) | Huang Teoh Zhang (2014 *TAR*) | Davis Tama-Sweet (2012 *CAR*)
- Proxy statements
  Mukhopadhyay Shivakumar (2015 *SSRN*)
- Conference calls
  Larcker Zakolyukina (2012 *JAR*) | Hobson Mayew Venkatachalam (2012 *JAR*) | Davis Ge Matsumoto Zhang (2015 *RASt*) | Brochet Naranjo Yu (2015 *SSRN*)
- IPO prospectuses
  Weiss-Hanley Hoberg (2010 *RFS*)

External sources

- Analyst reports
  Huang Zang Zheng (2014 *TAR*) | De Franco Hope Vyas Zhou (2015 *CAR*)
- News stories
  WSJ "Abreast of the Market" column: Tetlock (2007 *JoF*) | Dougal Engelberg Garcia Parsons (2012 *RFS*) | **WSJ and/or Dow Jones News Service articles:** Tetlock Saar Tsechansky Macskassy (2008 *JoF*) | Drake Guest Twedt (2014 *TAR*) | **Other (e.g. Factiva news source database:** Core Guay Larcker (2008 *JFE*) Bushee Core Guay Hamm (2010 *JAR*)
- Twitter posts
  Bartov Faurel Mohanram (2015 *SSRN*) | Bollen Mao Zeng (2011 *JCompSci*) | Blankespoor Miller White (2014 *TAR*)
- Investor message boards
  Antweiler Frank (2004 *JoF*)

Multiple channels

Kothari Li Short (2009 *TAR*): Corporations, analysts, business press

B 1. Text attributes: An emerging landscape

- Key challenge: **Construct validity**
  - Defining appropriate empirical constructs/proxies to capture a given theoretical concept

B 1. Text attributes: Towards a taxonomy [1)]

Theoretical concepts	Examples of common empirical constructs
Readability (complexity [2)], understandability [3)])	▪ Gunning-Fog, Flesch Reading Ease, and Flesch-Kincaid indices ▪ File size, word count ▪ Measures of "plain English" writing style
Similarity (redundancy, staleness, comparability, consistency, boilerplate)	▪ Variants of Jaccard similarity coefficients based on the intersection of words, bigrams or higher-order *n*-grams across documents/texts ▪ Vector Space Model
Tone (sentiment, affect, optimism/pessimism)	▪ Relative frequency of pessimistic (vs optimistic) words, based on some word list/dictionary, e.g., the Harvard psychosocial dictionary ▪ Holding tone constant: Vivid vs pallid language
Deceit (lying, manipulation)	▪ Deceptive language ▪ Vocal markers of cognitive dissonance ▪ Disclosure that is 'abnormal' in comparison
Other (selected)	▪ Concrete vs abstract language ▪ Vivid vs pallid language ▪ Ethics-related terms

1) Incomplete; terms vary | **2)** Challenging to distinguish economic complexity from linguistic complexity (see backup) |
3) Readability ≠ understandability (e.g., Smith Taffler 1992 AAAJ)

Readability: "ability of individual investors and analysts to assimilate valuation-relevant information from a financial disclosure" [1)]

Three main **approaches**:

- **Fog** index
  - For example, Li (2008 *JAE*)
- Measures of "**Plain English**" writing style
  - Prevalence of common style weaknesses, e.g., measured using text editing software "*StyleWriter—The Plain English Editor*"
  - For example, Miller (2010 *TAR*)
- Text **quantity** (word count, file size)
  - For example, Li (2008 *JAE*) and Loughran McDonald (2014 *JoF*)

1) Loughran McDonald (2014 *JoF*: 1649)

Gunning-Fog Index [1] = $0.4\left[\left(\frac{\text{words}}{\text{sentences}}\right) + 100\left(\frac{\text{complex words}^{2)}}{\text{words}}\right)\right]$

- Original purpose: Selecting suitable texts for students/readers of different age classes and education levels
  - Fog Index = 12 indicates 12 years of schooling needed for understanding a text upon first reading
- Applied to 10-Ks [3], analyst reports [4], and media articles [5] etc.
- Suitable for assessing readability of *financial* texts? [6]
  - First component: Average sentence length
    - In financial texts, as "." does not necessarily mark the end of a sentence
      - Lists, tables of contents, headers, decimal signs, abbreviations
  - Second component: Complex words
    - Do you find the following words particularly hard to understand?
      - Corporation, agreement, management, telecommunications, liability

1) Similar: Flesch Index and Flesch-Kincaid Index | **2)** Complex words: > 2 syllables | **3)** Examples include Li (2008 *JAE*); Lawrence (2013 *JAE*); Miller (2010 *TAR*); Lehavy Li Merkley (2011 *TAR*); Biddle, Hilary Verdi (2009 *JAE*) | **4)** De Franco Hope Vyas Zhou (2015 *CAR*) | **5)** For example the WSJ "Abreast of the market" column: Dougal Engelberg Garcia Parsons (2012 *RFS*) | **6)** Loughran McDonald (2014 *JoF*); Jones and Shoemaker 1994 (*JAL*)

Attempts to measure **overlap** of two or more texts

- Intersections of *n*-grams across texts (Jaccard similarity coefficients)
- Vector Space Model (VSM)

Labels and **interpretations**

- **Positive/neutral** notion
  - Accounting „**consistency**“: Peterson Schmardebeck Wilks (2015 *TAR*)
  - Audit client “**similarity**”: Brown Knechel (2016 *JAR*)
- **Negative** notion
  - **Redundancy**: Cazier Pfeiffer (2015 *AccHor*)
  - **Staleness** of news: Tetlock (2011 *RFS*)
  - **Boilerplate**: Lang Stice-Lawrence (2015 *JAE*)
  - Lack of “**modifications**”: Brown Tucker (2011 *JAR*)

B 1. Text attributes: Tone

- **Tone** (sentiment, affect) capture the relative frequency of pessimistic and optimistic words in a narrative
- Made popular by Tetlock (2007 *JoF*) and Tetlock Saar-Tsechansky Macskassy (2008 *JoF*) in studies on the sentiment of news stories
- Tone measures reflect the underlying **word lists** [1)], or dictionaries, used to classify words or phrases as optimistic or pessimistic
  - **Harvard University General Inquirer IV-4 psychosocial dictionary**
    - Examples: Tetlock (2007 *JoF*); Tetlock Saar Tsechansky Macskassy (2008 *JoF*); Engelberg (2008 *WP*); Kothari Li Short (2009 *TAR*)
  - ***Diction* optimism/pessimism word lists (dictionsoftware.com)**
    - Examples: Davis Piger Sedor (2012 *CAR*); Davis Ge Matsumoto Zhang (2014 *RASt*)
  - **Combinations of different word lists**
    - Examples: Rogers Van Buskirk Zechman (2011 *TAR*); Li (2010 *JAR*); Davis Tama-Sweet (2012 *CAR*)

1) Other word lists have been developed especially for financial texts, e.g., by Henry (2008 *JBusComm*) and Loughran and McDonald (2011 *JoF*)

B 1. Text attributes: Tone (cont'd) 1)

"Optimistic" *Diction* words common in financial texts

- outstanding
- respect
- determined

- power
- trust
- security
- authority

Examples of these words in non-optimistic contexts

- "common shares outstanding"
- "… with respect to …"
- "discount rate is determined by …"
- "electric power generation"
- "Contractual Trust Arrangement"
- "asset-backed security"
- "fiscal authority"

83% of the most common "optimistic" *Diction* words do not appear in Loughran McDonald's (2011 *JoF*) list of positive words custom-made for financial texts.

1) For this slide and the next, see Loughran McDonald (2015 *JBehavFin*)

"Pessimistic" *Diction* words common in financial texts	**Examples of these words in non-pessimistic contexts**
▪ not	▪ "EBITDA before special items is not defined in IFRS"
▪ no	▪ "sales growth shows no sign of slowing"
▪ gross	▪ "gross profit"
▪ lynch	▪ "Merrill Lynch"
▪ death	▪ "employee fluctuation includes terminations, retirements and deaths"
	▪ "plaintiffs allege bleeding and death"

70% of the most common "pessimistic" Diction words do not appear in Loughran McDonald's (2011 JoF) list of negative words custom-made for financial texts.

B 1. Text attributes: Deceit

- **Objective:** Detect cues for lying in firms' communications that can complement other known 'red flags' in detecting and predicting manipulation, accounting fraud, and financial misreporting

- **Example:** Deceptive language in earnings conference calls
  - Larcker Zakolyukina (2012 *JAR*)
  - Based on four theoretical perspectives on behavior during lying/deceit based on theory developed in Vrij (2008 *Detecting Lies and Deceit: Pitfalls and Opportunities*): Emotions, cognitive effort, control, and lack of embracement

My personal favorite: Enron conference call of April 17, 2001

"Operator: Richard Grubman of Highfield Capital (a hedge fund)

…

Richard Grubman: You're the only financial institution that cannot produce a balance sheet or cash flow statement with their earnings.

Jeff Skilling (Enron CFO): Thank you very much, we appreciate that.

Grubman: We appreciate that.

Skilling: A%%-hole."

B 2. Disclosure scores

- Using ATA methods to condense large bodies of text documents into ‚objective' proxies for the quality and/or quantity of disclosure

- Example of a relatively **abstract** proxy
  - Cooper He Plumlee (2015 *SSRN*)
  - Extract ***8K_Vdisc***, a measure of corporate voluntary disclosure [1)] from 8-Ks filed with the SEC
  - ***8K_Vdisc*** = count of *voluntary* reportable items for a firm within a calendar quarter, regardless of the number of 8Ks filed

- Example of a more **content-related** proxy
  - Grüning (2011 *EAR*)
  - *Artificial Intelligence Measure of Disclosure* (AIMD) captures the extent of corporate disclosure in English-language documents along 10 information dimensions

1) Note how the authors avoid referring to ***8K_Vdisc*** as a measure of voluntary disclosure *quality*. We believe not doing so is entirely adequate.

- **Idea of topic extraction:** Using ATA to identify (and extract) specific information from large text corpora to build variables of interest

- **Example 1**
  - Campbell Chen Dhaliwal Lu Steele (2014 *RASt*)
  - Extract quantitative measures of risk factor disclosures from 10-K filings

- **Example 2**
  - Daske Bischof Sextroh (2014 *JBFA*)
  - Identify fair value-related text passages in conference call Q&A sessions

- **Example 3**
  - Lundholm Li Minnis (2013 *JAR*)
  - Extract scaled number of references to competition in 10-K filings as a measure of a firm's competitive environment

Agenda

1. Introduction
2. Applications
3. Evidence
4. Methods
5. Critical discussion
6. Conclusion

Overview

I Disclosure attributes: Determinants

II Disclosure attributes = signals or predictors

Disclosure attributes

Tone	Deceptive language	Readability
↓	↓	↓
Stock market outcomes	Subsequent financial restatements	Subsequent performance / type of news

Economic outcomes

III Disclosure attributes = moderators or mediators

Fundamental signal

↓

← *Readability/complexity*

← *Disclosure tone*

← *Media/press coverage*

Market outcomes

I Determinants: Are text attributes chosen strategically?

- **Readability**: Li (2008 *JAE*)
  - Interested in „management obfuscation hypothesis“
  - 10-Ks of poor-performing firms less readable/more complex (effect statistically, *but not* economically, significant)
  - Profits of firms with more readable 10-Ks are more persistent (effect statistically *and* economically significant)
  - Concludes: “managers may be opportunistically structuring the annual reports to hide adverse information from investors”

- **Tone**: Huang Teoh Zhang (2014 *TAR*)
  - Estimate “abnormal positive tone” as a measure of discretionary tone
  - Positively associated with upward perception management, such as just meeting/beating thresholds, or future earnings restatements
  - Conclude: “evidence is consistent with managers using strategic tone management to mislead investors about firm fundamentals”

Does tone have information content?

- Tetlock (2007 *JoF*)
  - Negative tone in “Abreast of the Market” column (WSJ)
  - Predicts downward pressure on stock prices, reversion to fundamentals
  - Unusually high or low pessimism predicts high market trading volume
- Tetlock Saar-Tsechansky Macskassy (2008 *JoF*)
  - Negative tone in *WSJ* and *Dow Jones News Service* (*DJNS*) stories
  - Conveys negative information about future earnings above and beyond analysts’ forecasts and historical accounting data
  - Stock prices respond to the information in tone with a one-day delay
- Kothari Li Short (2009 *TAR*)
  - Analyze disclosure by firms, analysts, and the business press
  - Favorable (unfavorable) disclosures are accompanied by significant decreases (increases) in risk measures

Do text attributes help detect fraud?

- Larcker Zakolyukina (2012 *JAR*)
- Deceptive language in earnings conference calls
- Fraud indicated by:
  - More references to general knowledge ("you know")
  - More first-person plural pronouns ("we")
  - More general statements (e.g., "everybody", "anybody")
  - Extreme positive emotional words ("fantastic")
  - More tentative ("possibly"), less certain words ("definitely")
  - Negative statements (negation, anxiety, swear words, and anger)

II Disclosure attributes as signals or predictors (cont'd)

- Tone of earnings press releases [1)], earnings conference calls [2)], and MD&A disclosures [3)] has information content that triggers market reactions and helps predict future performance
- Compensation disclosures affect say on pay voting [4)]
- Greater target 10-K length enhances M&A efficiency [5)]
- Unusually optimistic disclosure tone in earnings announcements enhances shareholder litigation risk [6)]
- Frequency of 184 constraining words (e.g., *required, obligations, requirements, permitted, comply,* and *imposed*) has incremental diagnostic power for financial constraints [7)]

1) Davis Piger Sedor (2012 *CAR*) | **2)** Price Doran Peterson Bliss (2012 *JBF*) | **3)** Li (2010 *JAR*) | **5)** Kothari Li Short (2009 *TAR*) | **4)** Mukhopadhyay Shivakumar (2015 *SSRN*) | **5)** Chircop Tarsalewska (2015 *SSRN*) | **6)** Rogers Van Buskirk Zechman (2011 *TAR*) | **7)** Bodnaruk Loughran McDonald (2015 *JFQA*)

III Disclosure attributes as moderators or mediators

- Disclosure **complexity** can affect investors‘ trading behavior
  - Underreactions to 10-K filings increase in 10-K complexity, indicating complexity hampers the processing of information contained in 10-Ks [1)]
  - Small investors‘ trading around filings decreases in 10-K complexity [2)]
  - More readable disclosures yield stronger reactions from small investors [3)]
    - Interestingly, less so when potential readability differences are pointed out
- Two competing interpretations
  - <u>Information processing costs explanation</u>
    - „consistent with … more complex filings being too costly for small investors to process in the short window surrounding the filing date.” [4)]
  - <u>Processing fluency explanation</u>
    - „processing fluency from a more readable disclosure acts as a subconscious heuristic cue and increases investors’ beliefs that they can rely on the disclosure”, even if amount of information acquired is constant. [5)]

1) You and Zhang (2009 *RASt*) | **2)** Miller (2010 *TAR*) | **3)** Rennekamp (2012 *JAR*) | **4)** Miller (2010 *TAR*) | **5)** Rennekamp (2012 *JAR*); Elliott Rennekamp White (2015 *RASt*) show a similar positive effect for concrete (versus abstract) language, which „increases investors’ feelings of comfort in their ability to evaluate an investment.”

III Processing fluency – some examples [1)]

- Processing fluency is subjective, and represents how easy it feels to process information.
- Individuals like messages that feel easy to process.
- Processing fluency is affected by, for example,
  - Font: Easy to read – *hard to read*
  - Font size: Easy to read – hard to read
  - Color: Easy to read – hard to read
  - Rhyming: "What sobriety conceals, alcohol *reveals* / *unmasks*"
  - Simple versus complex synonyms: "deterioration" vs „decline"

- Processing fluency has been associated with higher ratings of truth, preference for the message and the messenger, willingness to rely on information, and confidence in judgments.

1) Taken, sometimes literally, from Rennekamp (2012 *JAR*: 1325-26)

10-K tone moderates market reactions to earnings announcements

- Henry Leone (2016 *TAR*)
  - Tone positively associated with market reaction to earnings announcements
  - Tone intensifies post-earnings announcement drift
- Similar results obtain for tone *change* in the MD&A of 10-Ks/10-Qs
  - Feldman Govindaraj Livnat Segal (2010 *RASt*)

Agenda

1. Introduction
2. Applications
3. Evidence
4. Methods
5. Critical discussion
6. Conclusion

- The field of natural language processing is like the field of econometrics:
  - vast
  - with a rich body of theory
  - full of philosophical details, disputes, and nuances
- We viewed our task as being similar to explaining OLS while breezing over the underlying theory and proofs.
- As you would with practical problems of inference with OLS (bias, collinearity, precision, etc.), we will focus on practical issues when using ATA

Language is highly ambiguous

- "You shall know a word by the company it keeps" (Firth 1957, p. 11)
  - E.g., the word "race"
- What are common patterns that occur in a language?
- "The major tool we use to identify those patterns is to count things, otherwise known as statistics…" (Manning Schütze 1999, p. 4)
- Statistical methods deal better with the ambiguity in natural languages (Manning Schütze 1999)
- Provides large sample methods to search and learn from word patterns.

Transforming text to numbers: Thinking like an accountant

"Accounting is the language of business."

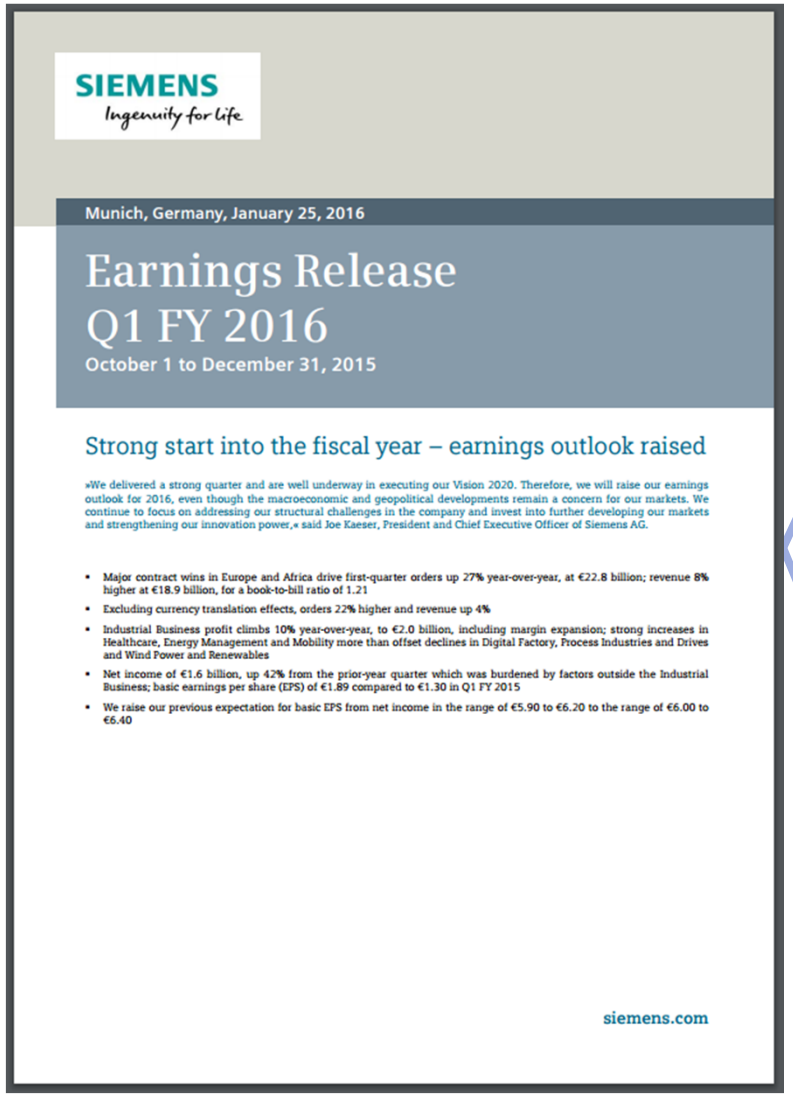
Accountants

- Assumes that transactions and their amounts characterize a firm / national economy.
- By properly aggregating transactions into accounts and relating them to each other as well as other firms, we conduct financial analysis, etc. to measure different attributes of a firm (performance, financial health etc.)
- Measurement problems (e.g., economic performance vs. accounting performance)

Text Analysts

- Assumes that terms and their frequencies characterize a text
- By properly aggregating term-frequencies and relating them to each other and other texts, one can potentially measure text attributes (sentiment, topic, readability, etc.)
- Measurement problems (e.g., readability vs. FOG Index)

Two areas of application in accounting research (so far)



SIEMENS
Ingenuity for life

Munich, Germany, January 25, 2016

Earnings Release
Q1 FY 2016
October 1 to December 31, 2015

Strong start into the fiscal year – earnings outlook raised

»We delivered a strong quarter and are well underway in executing our Vision 2020. Therefore, we will raise our earnings outlook for 2016, even though the macroeconomic and geopolitical developments remain a concern for our markets. We continue to focus on addressing our structural challenges in the company and invest into further developing our markets and strengthening our innovation power,« said Joe Kaeser, President and Chief Executive Officer of Siemens AG.

- Major contract wins in Europe and Africa drive first-quarter orders up 27% year-over-year, at €22.8 billion; revenue 8% higher at €18.9 billion, for a book-to-bill ratio of 1.21
- Excluding currency translation effects, orders 22% higher and revenue up 4%
- Industrial Business profit climbs 10% year-over-year, to €2.0 billion, including margin expansion; strong increases in Healthcare, Energy Management and Mobility more than offset declines in Digital Factory, Process Industries and Drives and Wind Power and Renewables
- Net income of €1.6 billion, up 42% from the prior-year quarter which was burdened by factors outside the Industrial Business; basic earnings per share (EPS) of €1.89 compared to €1.30 in Q1 FY 2015
- We raise our previous expectation for basic EPS from net income in the range of €5.90 to €6.20 to the range of €6.00 to €6.40

siemens.com

- Information Extraction
  - Disclosure scores
  - Novel datasets
  - Extract certain entities from (unstructured) data
  - E.g., firm names, audit partner, management forecasts, etc.

- Text Attribute Computation
  - Quantify certain properties/attributes of a given text numerically
  - E.g.,: sentiment, readability, complexity, similarity

General approaches to information extraction

Rules based

- E.g., Gate Software for Named Entity Recognition
- Simplest rule: keyword lists
- Precision varies a lot with tangibility of the desired entity/information and inherent structure of the texts.

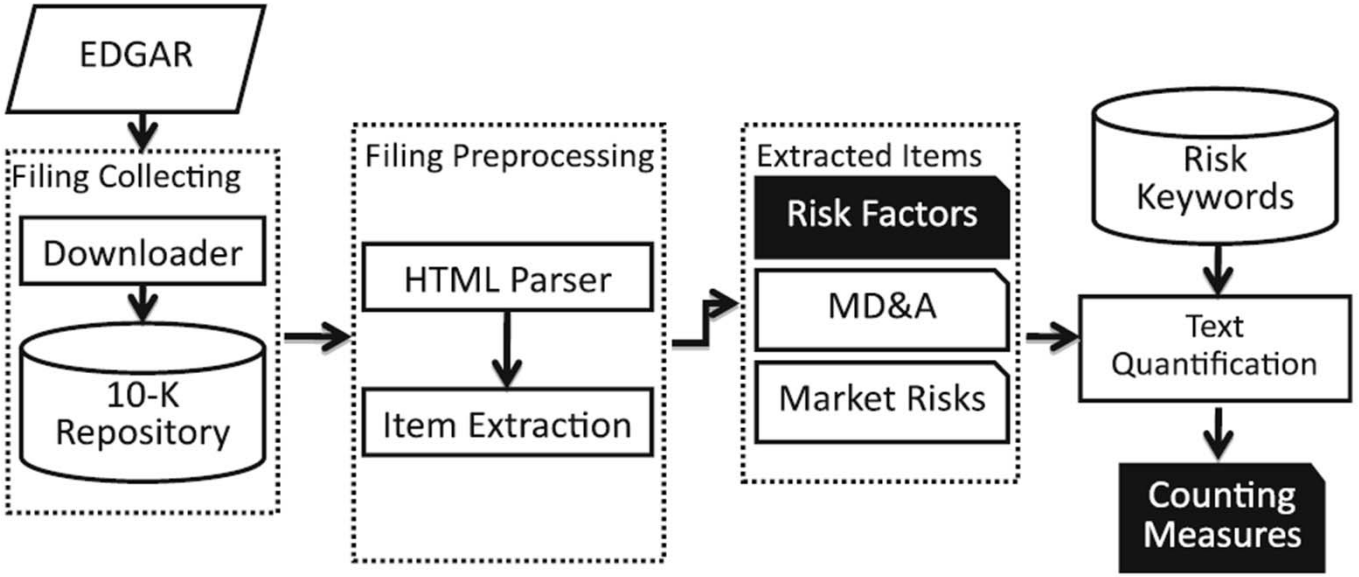
Unsupervised Machine Learning

- E.g., Huang, A., Lehavy, R., Zang, A., & Zheng, R. (2014). "Analyst information discovery and information interpretation roles: A topic modeling approach.", WP
- Essentially sophisticated clustering methods. Searches for term-cluster pattern that repeat across texts.
- Clusters need to be interpret by researchers afterwards.

Supervised Machine Learning

- "Trains" a model based on hand-coded training samples.
- Trained model is used to classify or predict new observations.
- Model tries to find the „best rules" for prediction
- Most widely used form in industrial applications today

Application example: extracting risk factors



Source: "Fig. 1 Analysis steps in constructing qualitative measures of risk factor disclosure for each company", Campbell, J. L., Chen, H., Dhaliwal, D. S., Lu, H. M., & Steele, L. B. (2014). The information content of mandatory risk factor disclosures in corporate filings. Review of Accounting Studies, 19(1), 396-455., p. 406

Extracted data and computed three content analysis measures per subsection:

1. total word count
2. total key word count (predefined keyword list (literature and clustering)
3. key word count by risk subcategory (predefined by literature)

Information extraction and classifiers – possible uses

- **Using term frequencies to classify documents.** For example:
  - Does an earnings release contain Non-GAAP pro-forma items?
  - Business versus non-business related news

- **Finding discerning key terms.** For example, what terms are associated with/predict:
  - Risk factors
  - Restatements
  - Qualified audit opinions
  - Management forecasts

Text attributes

- Often we are interested in certain attributes of a text (aka of a document, paragraph, sentence, etc.)
- Examples
  - **Similarity**: How similar are two texts? (For example, Tetlock 2011 uses text similarity to identify „old“ news.)
  - **Readability**: How readable is a text?
  - **Tone**: What sentiment does a text express?

- In each case, a text needs to be made machine readable in some form for further quantification

“The major tool we use to identify those patterns is **to count things**” (Manning Schütze 1999, p.4)

Transforming text to numbers in order to do statistics

Idea: Frequency of certain terms will be high or low, depending on tone, readability, topic, etc.

Identify grammatical structure often necessary for identifying terms

Count how often a "term" (depending on application: word, phrase, etc.) appears in a text

Transform text into a pattern of terms and their frequencies in the text.

Disclaimer: There are quite a few non-trivial issues involved in such a transformation. For now we abstract from these issues on purpose, but will return to some later (e.g.,. "good" and "not good" are obviously different "terms" for sentiment detection)

Frequency counts: Raw data for ATA applications

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Languages
العربية
Български
Čeština

Article Talk

International Financial Reporting Standards

From Wikipedia, the free encyclopedia
(Redirected from IFRS)

International Financial Reporting Standards (**IFRS**) are designed as a common global la
company accounts are understandable and comparable across international boundaries. Th
international shareholding and trade and are particularly important for companies that have
progressively replacing the many different national accounting standards. They are the rule
maintain books of accounts which are comparable, understandable, reliable and relevant as

IFRS, with the exception of *IAS 29 Financial Reporting in Hyperinflationary Economies* and
Approach under IAS 29, are authorized in terms of the historical cost paradigm. IAS 29 and
units of constant purchasing power paradigm.[1][2]

IFRS began as an attempt to harmonize accounting across the European Union but the valu
concept attractive around the world. However, it has been debated whether or not de facto
that were issued by IASC (the predecessor of IASB) and are still within use today go by the
Standards (IAS), while standards issued by IASB are called IFRS. IAS were issued betwee
International Accounting Standards Committee (IASC). On 1 April 2001, the new Internation
took over from the IASC the responsibility for setting International Accounting Standards. Du
adopted existing IAS and Standing Interpretations Committee standards (SICs). The IASB h
Reporting Standards".

> *In the absence of a Standard or an Interpretation that specifically applies to a transac*
> *accounting policy that results in information that is relevant and reliable. In making tha*
> *recognition criteria, and measurement concepts for assets, liabilities, income, and exp*

Criticisms of IFRS are (1) that they are not being adopted in the US (see GAAP), (2) a numb
Hyperinflationary Economies had no positive effect at all during 6 years in Zimbabwe's hype
offered no response to the last criticism while IAS 29 is currently (March 2014) being implem

Contents [hide]

Source: https://en.wikipedia.org/wiki/International_Financial_Reporting_Standards

20 most frequent terms:

Term	Frequency
the	383
of	260
in	137
and	124
to	119
ifrs	95
financial	80
is	73
for	72
companies	55
that	54
are	53
standards	53
accounting	50
capital	49
as	42
statements	37
not	35
be	35
or	33

Length of unfiltered vocabulary: 1166

Adding structure to identify and classify term patterns

Identify sentence structure using **"part-of-speech"** tags (POS tags):*

"International Financial Reporting Standards (IFRS) are designed as a common global language for business affairs so that company accounts are understandable and comparable across international boundaries."

↓

('International', 'NNP') ('Financial', 'NNP') ('Reporting', 'NNP') ('Standards', 'NNP') ('(', '(') ('IFRS', 'NNP') (')', ')') ('are', 'VBP') ('designed', 'VBN') ('as', 'IN') ('a', 'DT') ('common', 'JJ') ('global', 'JJ') ('language', 'NN') ('for', 'IN') ('business', 'NN') ('affairs', 'NNS') ('so', 'IN') ('that', 'DT') ('company', 'NN') ('accounts', 'NNS') ('are', 'VBP') ('understandable', 'JJ') ('and', 'CC') ('comparable', 'JJ') ('across', 'IN') ('international', 'JJ') ('boundaries', 'NNS') ('.', '.')

- e.g., "*International Financial Reporting Standards (IFRS)*" is labeled as a sequence of "NNP" tags–a proper noun.
  - Can be used to identify it as one, and not 5 separate terms
- Can be used to find negations ("not … good"), etc.

* Using the UPenn tagset here (e.g., JJ = adjective or numeral, ordinal)

Why POS tagging?

- Automated POS taggers trained on (hopefully) representative texts, but not without error
- Essential for important preprocessing tasks such as:
  - Assessing word sense (e.g., are we talking about “race” in the sense of a contest?)
  - Identifying terms consisting of many words (e.g., “financial statements”)
  - Identifying negations (“not … good”) for sentiment analysis and similar applications
  - Stemming (count swimming, swam, swim as one term where appropriate)

Text attribute 1: Sentiment

"We delivered a strong quarter and are well underway in executing our Vision 2020. Therefore, we will raise our earnings outlook for 2016, even though the macroeconomic and geopolitical developments remain a concern for our markets. We continue to focus on addressing our structural challenges in the company and invest into further developing our markets and strengthening our innovation power…"

Q1/2016 Earnings Release of Siemens AG

Transform
(Term = combination of words + POS tag*)

Word	Tag	#
Therefore	RB	1
Vision	NN	1
We	PRP	2
a	DT	2
addressing	VBG	1
and	CC	4
are	VBP	1
challenges	NNS	1
company	NN	1
concern	NN	1
continue	VBP	1
delivered	VBD	1
developing	VBG	1
…	…	…
power	NN	1
quarter	NN	1
raise	VB	1
remain	VBP	1
strengthening	VBG	1
strong	JJ	1
structural	JJ	1
the	DT	2
though	IN	1
to	TO	1
underway	RB	1
we	PRP	1
well	RB	1
will	MD	1

* Using the UPenn tagset here (e.g., JJ = adjective or numeral, ordinal)

Text attribute 1: Sentiment (cont'd)

"We delivered a strong quarter and are well underway in executing our Vision 2020. Therefore, we will raise our earnings outlook for 2016, even though the macroeconomic and geopolitical developments remain a concern for our markets. We continue to focus on addressing our structural challenges in the company and invest into further developing our markets and strengthening our innovation power…"

Q1/2016 Earnings Release of Siemens AG

Transform
(Term = combination of words + POS tag*)

Word	Tag	#
Therefore	RB	1
Vision	NN	1
We	PRP	2
a	DT	2
addressing	VBG	1
and	CC	4
are	VBP	1
challenges	NNS	1
company	NN	1
concern	NN	1
continue	VBP	1
delivered	VBD	1
developing	VBG	1
…	…	…
power	NN	1
quarter	NN	1
raise	VB	1
remain	VBP	1
strengthening	VBG	1
strong	JJ	1
structural	JJ	1
the	DT	2
though	IN	1
to	TO	1
underway	RB	1
we	PRP	1
well	RB	1
will	MD	1

…?

Now what?

* Using the UPenn tagset here (e.g., JJ = adjective or numeral, ordinal)

Text attribute 1: Sentiment (cont'd)

- Can we somehow identify which terms convey sentiment?
- **Approach 1**: Computer-derived classification
  - Collect a hand-coded training sample
  - Label training sample documents as positive/negative: 1/0
  - Run/test a classification model (e.g., logistic regression) with all terms (word plus position in sentence) as variables
  - Model parameters tell you which terms have the most discriminatory power in this sample.
  - Extract a list of most discriminating positive/negative words, or use model directly for prediction
  - Especially useful if sentiment contextual (e.g., Twitter sentiment)
- **Approach 2**: Human-generated classification
  - Example: Harvard General Inquirer
  - Widely used in accounting research

Text attribute 1: Sentiment (cont'd)

	Entry	Positiv	Negativ	...184 other classes ...	Othtags	Defined
1	A				DET ART	...
2	ABANDON		Negativ		SUPV	
3	ABANDONMENT		Negativ		Noun	
...						
213	ADVANCE#1	Positiv			SUPV	47% verb: To move or bring forward, improve, promote
214	ADVANCE#1	Positiv			Noun	20% noun-adj: A moving forward, improvement, approach, in front, prior
215	ADVANCE#1	Positiv			Modif	20% adj: "Advanced"--forward, in front, progressive
216	ADVANCE#1				LY	13% idiom-adv: "In advance"--before, beforehand
...						

Source: Excerpt of the Harvard General Inquirer spreadsheet file (http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm)

Harvard General Inquirer

"The General Inquirer is basically a mapping tool."

- Maps each text file with counts on dictionary-supplied categories.
- Most of processing time spent on identifying commonly used word senses. (e.g., '"race" as a contest, "race" as moving rapidly, "race" as a group of people of common descent').
- Applies word stemming: Swimming -> swim
- "The 182 General Inquirer categories were developed for social-science content-analysis research applications, not for text archiving, automatic text routing, automatic text classifying, or other natural-language processing objectives... People, not computers, created these categories, although some category developers drew upon cluster analyses produced by computers. Many categories were initially created to represent social-science concepts of several grand theories that were prominent at the time the system was first developed..."

Source: paraphrased from http://www.wjh.harvard.edu/~inquirer/3JMoreInfo.html

Text attribute 1: Sentiment (cont'd)

Term	Tag	#
Therefore	RB	1
Vision	NN	1
We	PRP	2
a	DT	2
addressing	VBG	1
and	CC	4
are	VBP	1
challenges	NNS	1
company	NN	1
concern	NN	1
continue	VBP	1
delivered	VBD	1
developing	VBG	1
…	…	…
power	NN	1
quarter	NN	1
raise	VB	1
remain	VBP	1
strengthening	VBG	1
strong	JJ	1
structural	JJ	1
the	DT	2
though	IN	1
to	TO	1
underway	RB	1
we	PRP	1
well	RB	1
will	MD	1

Transform, word sense disambiguation, stemming, etc

Term	Tag	#	
Therefore	RB	1	
Vision	NN	1	
We	PRP	2	
a	DT	2	
addressing	VBG	1	
and	CC	4	
are	VBP	1	
challenges	NNS	1	Neg
company	NN	1	
concern	NN	1	Neg
continue	VBP	1	
delivered	VBD	1	
developing	VBG	1	
…	…	…	
power	NN	1	Pos
quarter	NN	1	
raise	VB	1	
remain	VBP	1	
strengthening	VBG	1	Pos
strong	JJ	1	Pos
structural	JJ	1	
the	DT	2	
though	IN	1	
to	TO	1	
underway	RB	1	
we	PRP	1	
well	RB	1	
will	MD	1	

Transform and compute

$$Neg = \frac{negative\ words}{total\ words}$$

Proxy for the degree of negative sentiment

Fog Index

- A weighted measure, scaled to give the number of years of education needed to read the text.
- Count of "complex" words a significant determinant. Especially in financial texts
- Professional programs label terms as complex if consisting of three or more syllables.
- But there are exceptions (depending on how readability is defined in context)
  - Don’t count proper nouns (e.g., “International Financial Reporting Standards”),
  - Omit familiar jargon,
  - Sometimes split compound words (e.g., “skyscraper”).
  - Ignore common suffixes (such as -es, -ed, or -ing) as a syllables (e.g., reporting)

Text attribute 2: Readability (cont'd)

"International Financial Reporting Standards (**IFRS**) are designed as a common global language for business affairs so that company accounts are understandable and comparable across international boundaries"

Term	Freq
are	2
international	2
accounts	1
financial	1
that	1
standards	1
so	1
reporting	1
language	1
ifrs	1
global	1
for	1
designed	1
across	1
comparable	1
company	1
common	1
business	1
boundaries	1
as	1
and	1
affairs	1
understandable	1

Filter →

Term	Freq
are	2
International	2
accounts	1
financial	1
that	1
standards	1
so	1
reporting	1
language	1
ifrs	1
global	1
for	1
designed	1
across	1
comparable	1
company	1
common	1
business	1
boundaries	1
as	1
and	1
affairs	1
understandable	1

Simplified definition:
Complex = 3+ Syllables

Fog Index

$$0.4\left(\frac{words}{sentences} + 100\frac{complex\ words}{words}\right)$$

0.4 (25 /1+ 100*11/25) = 27.6

Transform

Text attribute 2: Readability (cont'd)

"International Financial Reporting Standards (**IFRS**) are designed as a common global language for business affairs so that company accounts are understandable and comparable across international boundaries."

Term	Freq
are	2
International financial reporting standards (ifrs)	1
international	1
accounts	1
that	1
so	1
language	1
global	1
for	1
designed	1
across	1
comparable	1
company	1
common	1
business	1
boundaries	1
as	1
and	1
affairs	1
understandable	1

Filter →

Term	Freq
are	2
International financial reporting standards (ifrs)	1
international	1
accounts	1
that	1
so	1
language	1
global	1
for	1
designed	1
across	1
comparable	1
company	1
common	1
business	1
boundaries	1
as	1
and	1
affairs	1
understandable	1

Fog Index

$$0.4\left(\frac{words}{sentences} + 100\frac{complex\ words}{words}\right)$$

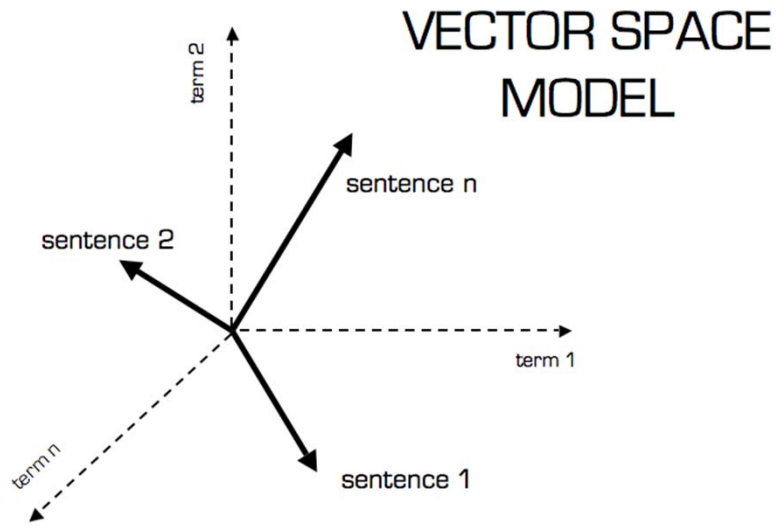
0.4 (25 /1+ 100*7/25) = 21.2

Transform

Text attribute 3: Similarity

- X = "This is a text."
- Y = "This is a text. This is a text."
- Z = "This is a more elaborate text containing more words“

Term	X	Y	Z
a	1	2	1
containing	0	0	1
elaborate	0	0	1
is	1	2	1
more	0	0	2
text	1	2	1
this	1	2	1
words	0	0	1



Compute angle between vectors:
Similarity(X,Y): {'cos': 1.00, ‘Angle': 0.0}
Similarity(X,Z): {'cos': 0.55, ‘Angle': 56.8}
Similarity(Y,Z): {'cos': 0.55, ‘Angle': 56.8}

Picture source: http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/

Text attribute 3: Similarity (cont'd)

- Textual similarity of documents, paragraphs, passages, etc.
- Used in older search engine applications and other information retrieval systems
- Used in accounting and finance research to:
  - Identify old news, redundant text
  - "Boilerplate" text
- Other commonly used similarity measure: Jaccard distance
  - Measures how many common elements two sets contain
  - Ratio of the intersection of set elements to the union of set elements.
  - E.g., as a measure of stale news (Tetlock 2011)

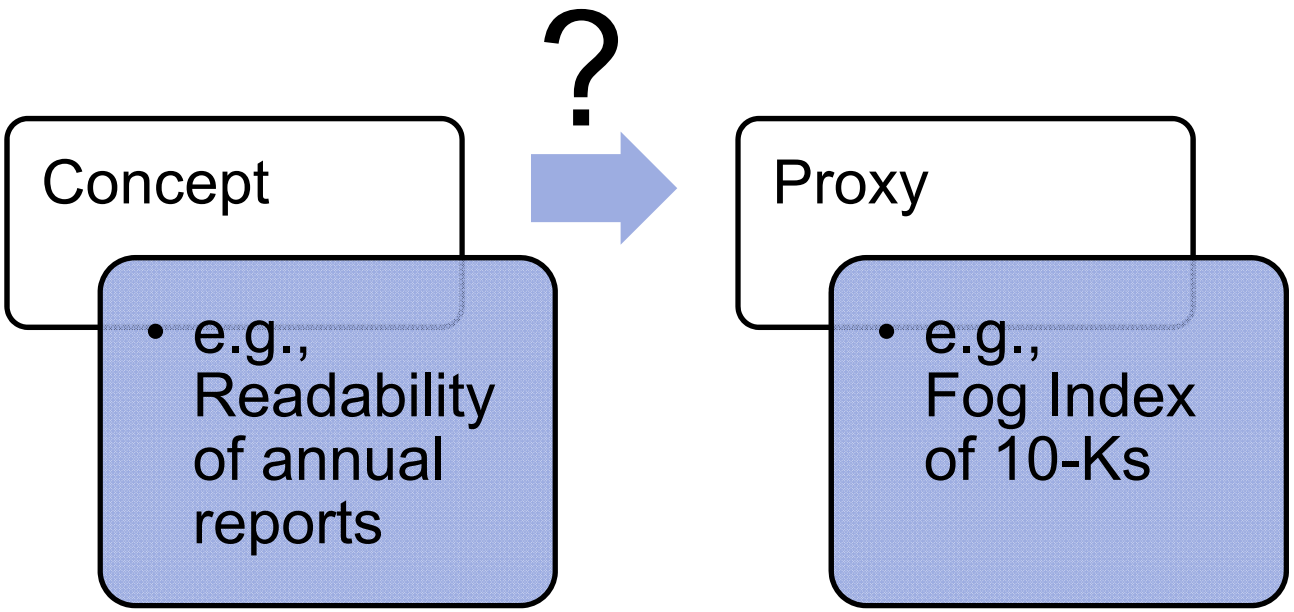
Agenda

1. Introduction
2. Applications
3. Evidence
4. Methods
5. Critical discussion
6. Conclusion

Are these approaches any good?

After this brief overview, we need to settle a few things

- How valid are these proxies?
- How can we put structure on this question?
- Example:



A measurement error perspective

- We are mostly concerned about correct inferences in regressions.
- So, let‘s frame the problem as a form of **measurement error**:

$$Proxy = Concept^* + MErr$$

- For example:

$$Fog_{10K} = Readability^*_{AR} + MErr$$

Problems with constructs

True relation:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 Readabilty_{AR}^{*} + \epsilon$$

What we measure:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 Fog_{10K} + (\epsilon - \alpha_2 MErr)$$

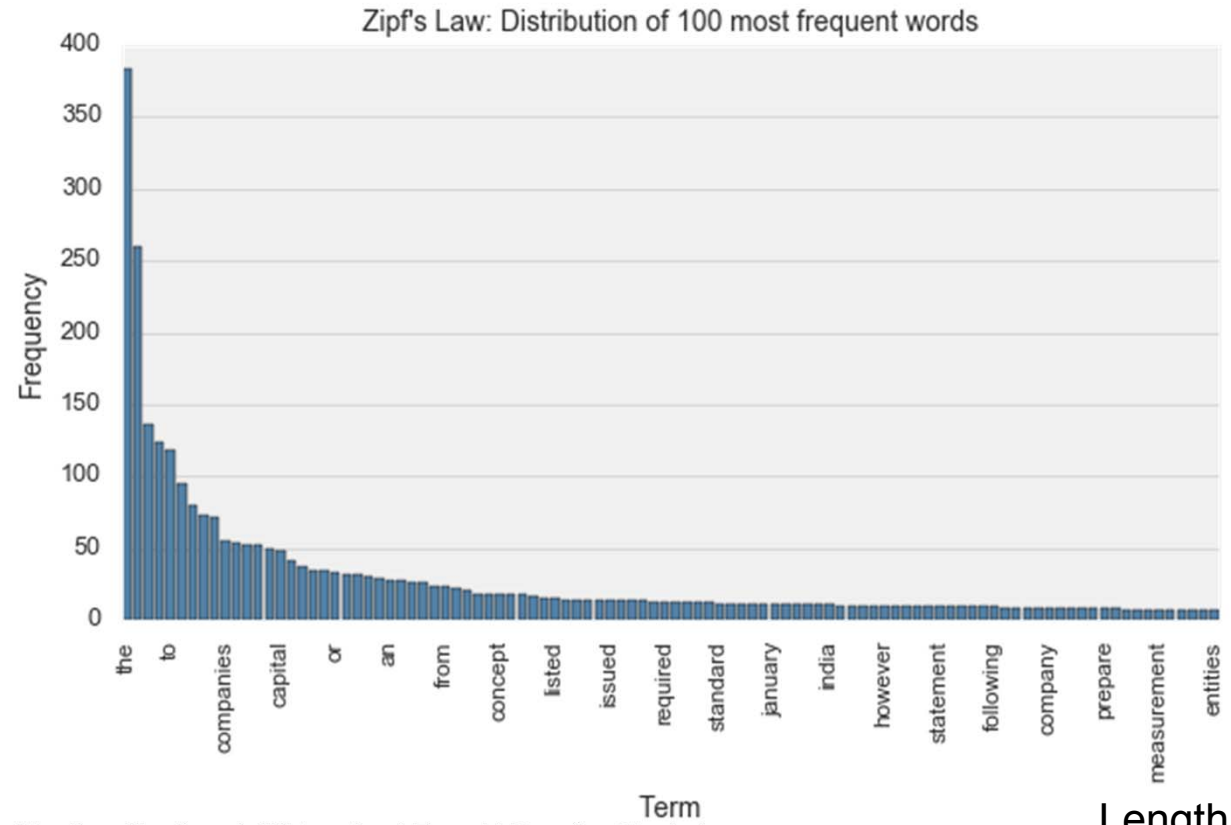
- If $\sigma(Fog_{10K}, MErr) = 0$, "only" increases standard errors.
- **Problem** if $\sigma(Fog_{10K}, MErr) \neq 0$.
  - For example, if Fog_{10K} measures annual report readability worse for texts with high Fog_{10K} than for texts with low Fog_{10K}.
  - Extreme case: $\hat{\alpha}_2$ will be zero if Fog_{10K} measures only noise.
  - Direction of bias tricky, if Fog_{10K} also correlated with omitted variables (V^*), i.e. picks up something unrelated to readability.

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 Fog_{10K} + (\epsilon - \alpha_2 MErr + \alpha_3 V^*)$$

Sources of measurement error 1: What is a term?

- Highly context specific and large leeway („financial statement“ versus „financial“ and „statement“
- Words can have multiple meanings (e.g., “race”)
- „run“, „ran“, „running“ one term? „company“, „companies“? (stemming or lemmatization methods).
  - Sometimes important to count terms properly
  - Importance depends on the language
- Most sophisticated programs:
  - Define „term“ as a word plus its position in the sentence (e.g., „absent“ could be a verb or adjective)
  - Try some kind of word sense disambiguation.

Sources of measurement error 2: Zipf's law



Source: https://en.wikipedia.org/wiki/International_Financial_Reporting_Standards

20 most frequent terms:

Term	Frequency
the	383
of	260
in	137
and	124
to	119
ifrs	95
financial	80
is	73
for	72
companies	55
that	54
are	53
standards	53
accounting	50
capital	49
as	42
statements	37
not	35
be	35
…	…

Length of unfiltered vocabulary: 1166

Sources of measurement error 2: Zipf's law (cont'd)

- Term-Frequency distribution in texts:
  - Zipf's Law: some words make up most of a text.
  - E.g., 52 complex words of ca. 45,000 complex words in 10-Ks make up ca. 25% of all complex words occurring in that sample (Loughran McDonald 2014, p. 1645)
  - Often need to filter for „stopwords" („and", „the", etc.) and other common non-informative but context specific words.
- Really make sure the most frequent words measure what you want.
- Variation in highly frequent terms usually dominates variation in the proxy (sentiment, Fog, etc.).
- If that is the case, even small measurement error (i.e. 2 out of 300 terms), when arising in highly frequent terms, often leads to severe, correlated measurement error.
- Consider weighting terms
  - Sdee, for example, Henry and Leone (2016 *TAR*)

Sources of measurement error 2: Zipf's law (cont'd)

"If the Combination completes, the existing Shell Shareholders and the former BG Shareholders will own a smaller percentage of Shell than they currently own of Shell and BG, respectively. Existing Shell Shareholders and former BG Shareholders will own approximately 81% and 19% respectively of the ***outstanding*** *Shell Shares. As a consequence, the number of voting rights which can be exercised and the influence which may be exerted by them in respect of the Combined Group will be reduced."* [1)]

1) Shell bid prospectus (22. Dec. 2015. p.25)

Example: Measuring tone

"If the Combination completes, the existing Shell Shareholders and the former BG Shareholders will own a smaller percentage of Shell than they currently own of Shell and BG, respectively. Existing Shell Shareholders and former BG Shareholders will own approximately 81% and 19% respectively of the ***outstanding*** *Shell Shares. As a consequence, the number of voting rights which can be exercised and the influence which may be exerted by them in respect of the Combined Group will be reduced."* [1)]

- outstanding (i.e. "excellent") vs. (shares) outstanding
- How frequent is "common shares outstanding"
  - … in an IPO prospectus?
  - … in M&A deal announcements?
- Very frequent in an IPO prospectus. Does not help explain variation? Consider reweighting it?
- In M&A deals, frequency maybe tied to whether share or cash deal.
  - Sentiment measure does not (only) pick up tone, but also deal type?

1) Shell bid prospectus (22. Dec. 2015. p.25)

Sources of measurement error 3: Representative text?

- Is representative text available for the question at hand?
- Not so much a problem with the method, but an important aspect for validity.
- Can affect the accuracy of the automated methods (like POS tagging)
- Example:
  - Twitter sentiment
  - What is the demographic that tweets, and why?
  - Is Twitter sentiment, even if appropriately capturing the sentiment of that demographic, always a good indicator for broader sentiment in the stock market? In the economy? Society?
  - As an aside, measuring sentiment on 140 character Twitter feeds posts very unique challenges.

Agenda

1. Introduction
2. Applications
3. Evidence
4. Methods
5. Critical discussion
6. Conclusion

ATA: a promising area to engage in?

Opportunities

- Researchers share word lists and code
- Extract innovative datasets to address questions of interest using large-sample approaches
- Create finer, more tailored measures of phenomena of interest

Challenges

- Correlated omitted variables
  - For example: Separating (intentional) linguistic complexity from underlying economic complexity
  - Measuring a causal effect of readability holding complexity fixed might not be possible (Loughran MacDonald 2014 JoF, p. 1646)
- Online filing requirements currently put US researchers at an advantage
- Crowding out of subtle, subjective, but potentially more accurate measures not readily captured by algorithms

Current trends

- Challenging initial interpretations and elucidating causal mechanisms
  - Disclosure readability and investor responses: Information processing costs or perceived reliability due to processing fluency? 1)
  - Disclosure tone and investor responses: Signaling of private information, manager traits or tone management? 2)
- New, more nuanced measures – endless possibilities, but do they all make sense?
- „Under-researched“ disclosures
- Trying to understand „who writes“

1) Essentially, You and Zhang (2009 *RASt*) and Miller (2010 *TAR*) versus Rennekamp (2012 *JAR*) and Elliott Rennekamp White (2015 *RASt*) |
2) Essentially, Davis Piger Sedor (2012 *CAR*) versus Davis Ge Matsumoto Zhang (2015 *RASt*) versus Huang Teoh Zhang (2014 *TAR*)

Summary

- Goal of this talk: Introduce developments in Automated Textual Analysis (ATA) and related methods; maintaining critical distance
- ATA is the examination (and retrieval) of document content using computer algorithms
- Increasingly relevant in practice and research
- ATA is being applied to a diverse range of documents
- To *measure* attributes of disclosure, and to *construct* disclosure proxies, novel variables and large datasets in a replicable way
- Facilitates research into (1) disclosure properties as fundamental signals and (2) as moderators of the effect of such signals
- ATA subject to challenges of construct validity and causal inference
- ATA holds great promise, but biggest potatoes have been gathered
- Construct validity and transparency key requirements for progress

Thank you

References

Allee, Kristin D./DeAngelis, Matthew D. (2015): The Structure of Voluntary Disclosure Narratives: Evidence from Tone Dispersion, Vol. 53, pp. 241-274.

Antweiler, Werner/Frank, Murray Z. (2004): Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards, in: Journal of Finance, Vol. LIX, pp. 1259-1294.

Bartov, Eli/Faurel, Lucile/Mohanram, Partha (2015): Can Twitter Help Predict Firm-Level Earnings and Stock Returns?

Beal Frazier, Katherine/Ingram, Robert W./ Mack Tennyson, B. (1984): A Methodology for the Analysis of Narrative Accounting Disclosures, in: Journal of Accounting Research, Vol. 22, pp. 318-331.

Biddle, Gary C./Hilary, Gilles/Verdi, Rodrigo S. (2009): How does financial reporting quality relate to investment efficiency?, in: Journal of Accounting and Economics, Vol. 48, pp. 112-131.

Bischof, Jannis/Daske, Holger/Sextroh, Christoph (2014): Fair Value-related Information in Analysts´ Decision Process: Evidence from the Financial Crisis, in: Journal of Business Finance and Accounting, Vol. 41, pp. 363-400.

Blankespoor, Elizabeth/ Miller, Gregory S./White, Hal D. (2014): The Role of Dissemination in Market Liquidity: Evidence from Firms´ Use of Twitter, in: The Accounting Review, Vol. 89, pp. 79-112.

Bodnaruk, Andriy/Loughran, Tim/McDonald, Bill (2015): Using 10-K Text to Gauge Financial Constraints, in: Journal of Finance and Quantitative Analysis, Vol. 50, pp. 623-646.

References (cont'd)

Bollen,Johan/Mao, Huina/Zeng, Xiao-Jun (2011): Twitter mood predicts the stock market, in: Journal of Computational Science, Vol. 2, pp. 1-8.

Bonsall, Samuel B. IV/ Leone, Andrew J./Miller, Brian P./Rennekamp, Kristina (2015): A Plain English Measure of Financial Reporting Readability.

Brochet, Francois/Naranjo, Patricia/Yu, Gwen (2015): The capital market consequences of language barriers in the conference calls of non-U.S. firms.

Brown, Stephen V./Knechel, W. Robert (2016): Auditor-Client Compatibility and Audit Firm Selection.

Brown, Stephen V./Tucker, Jennifer Wu (2011): Large-Sample Evidence on Firms' Year-over-Year MD&A Modifications, in: Journal of Accounting Research, Vol. 49, pp. 309-346.

Bushee, Brian J./Core, John E./Guay, Wayne/Hamm, Sophia J. W. (2010): The Role of the Business Press as an Information Intermediary, in: Journal of Accounting Research, Vol. 48, pp. 1-20.

Cazier, Richard A./Pfeiffer, Ray (2015): Why are 10-K Fillings so long?, in: Accounting Horizons.

Chen, Jason V./Li, Feng (2013): Estimating the Amount of Estimation in Accruals.

References (cont'd)

Chircop, Justin/Tarsalewska, Monika (2015): Annual Report Length and M&A Efficiency.

Cooper, Michael J./He, Jing/Plumlee, Marlene A. (2015): Voluntary Disclosure and Investor Sentiment.

Core, John E./Guay, Wayne/Larcker David F. (2008): The power of the pen and executive compensation, in: Journal of Financial Economics, Vol. 88, pp. 1-25.

Davis, Angela K./Piger, Jeremy M./Sedor, Lisa M. (2012): Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language, in: Contemporary Accounting Research, Vol. 29, pp. 845-868.

Davis, Angela K./Ge, Weili/ Matsumoto, Dawn/ Zhang, Jenny Li (2015): The effect of manager-specific optimism on the tone of earnings conference calls, in: Review of Accounting Studies, Vol. 20, pp. 639-673.

Davis, Angela K./Tama-Sweet, Isho (2012): Managers' Use of Language Across Alternative Disclosure Outlets: Earnings Press Releases versus MD&A, in: Contemporary Accounting Research, Vol. 29, pp. 804-837.

Dikolli, Shane S./Keusch, Thomas/Mayew, William J./Steffen, Thomas D. (2014): Using Shareholder Letters to Measure CEO Integrity.

Dougal, Casey/Engelberg, Joseph/Garcia, Diego/Parsons, Christopher A. (2012): Journalists and the Stock Market, in: The Review of Financial Studies, Vol. 25, pp. 639-680.

Drake, Michael S./Guest, Nicholas M./Twedt, Brady J. (2014): The Media Mispricing: The Role of the Business Press in the Pricing of Accounting Information, in: The Accounting Review, Vol. 89, pp. 1673-1701.

References (cont'd)

Engelberg, Joseph (2008): Costly Information Processing: Evidence from Earnings Announcements.

Elliott, W. Brooke/Rennekamp, Kristina/White, Brian J. (2015): Does concrete language in disclosures increase willingness to invest?, in: Review of Accounting Studies, Vol. 20, pp.839-865.

Feldman, Ronen/Govindaraj, Suresh/Livnat, Joshua/Segal, Benjamin (2010): Management´s tone change, post earnings announcement drift and accruals, in: Review of Accounting Studies, Vol. 15, pp. 915-953.

de Franco, Gus/Hope, Ole-Kristian/Vyas, Dushyantkumar/Zhou, Yibin (2015): Analyst Report Readability.

Grüning, Michael (2011): Artificial Intelligence Measurement of Disclosure (AIMD), in: European Accounting Review, Vol. 20, pp. 485-519.

Hales, Jeffrey/Kuang, Xi Jason/ Venkataraman, Shankar (2011): Who believes the Hype? An Experimental Examination of How Language Affects Investor Judgments, in: Journal of Accounting Research, Vol. 49, pp. 223-255.

Henry, Elaine (2008): Are investors influenced by how earnings press releases are written?, in: Journal of Business Communication, Vol. 45, pp. 363-407.

Henry, Elaine/Leone, Andrew J. (2016): Measuring Qualitative Information in Capital Markets Research: Comparison of Alternative Methodologies to Measure Disclosure Tone, in: The Accounting Review, Vol. 91, pp. 153-178.

Hoberg, Gerard/Lewis, Craig (2015): Do Fraudulent Firms Produce Abnormal Disclosure?

Hoberg, Gerard/Phillips, Gordon (2010): Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis, in: The Review of Financial Studies, Vol. 23, pp. 3773-3811.

References (cont'd)

Huang, Allen H./Zang, Amy Y./Zheng, Rong (2014): Evidence on the Information Content of Text in Analyst Reports.

Humpherys, Sean L./Moffitt, Kevin C./Burns, Mary B./Burgoon, Judee K./Felix, William F. (2011): Identification of fraudulent financial statements using linguistic credibility analysis, in: Decision Support Systems, Vol. 50, pp. 585-594.

Hussainey, Khaled/Schleicher, Thomas/Walker, Martin (2003): Untertaking large-scale disclosure studies when AIMR-FAF ratings are not available: the case of prices leading earnings, in: Accounting and Business Research, Vol. 33, pp. 275-294.

Hwang, Byoung-Hyoun/Kim Hugh Hoikwang (2015): It pays to write well.

Jegadeesh, Narasimhan/Wu, Di (2013): Word power: A new approach for content analysis, in. Journal of Finance Economics, Vol. 110, pp. 712-729.

Jones, Michael John/Shoemaker, Paul A. (1994): Accounting Narratives: A review of empirical studies of content and readability, in: Journal of Accounting Literature.

Lang, Mark/Stice-Lawrence, Lorien (2015): Textual Analysis and International Financial Reporting: Large Sample Evidence, in: Journal of Accounting and Economics.

Larcker, David F./Zakolyukina, Anastasia A. (2012): Detecting Deceptive Discussions in Conference Calls, in: Journal of Accounting Research, Vol. 50, pp. 495-540.

Lawrence, Alastair (2013): Individual investors and financial disclosure, in: Journal of Accounting and Economics, Vol. 56, pp. 130-147.

References (cont'd)

Lehavy, Reuven/Li, Feng/Merkley, Kenneth (2011): The Effect of Annual Report Readability on Analyst Following and the Properties of Their Earnings Forecasts, in: The Accounting Review, Vol. 86, pp. 1087-1115.

Li, Feng (2008): Annual report readability, current earnings, and earnings persistence, in: Journal of Accounting and Economics, Vol. 45, pp. 221-247.

Li, Feng (2010): The Information Content of Forward-Looking Statements in Corporate Filings – A Naïve Bayesian Machine Learning Approach, in: Journal of Accounting Research, Vol. 48, pp. 1049-1102.

Li, Feng/Lundholm, Russell/Minnis, Michael (2013): A Measure of Competition Based on 10-K Fillings, in: Journal of Accounting Research, Vol. 51, pp. 399-436.

Loughran, Tim/McDonald, Bill (2014): Measuring Readability in Financial Disclosures, in: Journal of Finance, Vol. LXIX, pp. 1643-1671.

Loughran, Tim/McDonald, Bill (2015): The Use of World Lists in Textual Analysis, in: Journal of Behavioral Finance.

Loughran, Tim/McDonald, Bill (2011): When Is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks, in: The Journal of Finance, Vol. LXVI, pp. 35-65.

Loughran, Tim/McDonald, Bill/Yun, Hayong (2009): A Wolf in Sheep´s Clothing: The Use of Ethics-Related Terms in 10-K Reports, in: Journal of Business Ethics, Vol. 89. 39-49.

Hobson, Jessen L./Mayew, William H./Venkatachalam, Mohan (2012): Analyzing Speech to Detect Financial Misreporting, in: Journal of Accounting Research, Vol. 50. 349-392.

References (cont'd)

McMullin, Jeff L. (2014): Can I Borrow your Footnotes? Learning and Network Benefits of Footnote Similarity.

Miller, Brian P. (2010): The Effects of Reporting Complexity on Small and Large Investor Trading, in: The Accounting Review, Vol. 85, pp. 2107-2143.

Mukhopadhyay, Tathagat/Shivakuma, Lakshmanan (2015): Do Compensation Disclosures Matter for SoP Voting?

Nelson, Karen K./Pritchard, Adam C. (2014): Carrot or Stick? The Shift from Voluntary to Mandatory Disclosure of Risk Factors.

Peterson, Kyle/Schmardebeck, Roy/Wilks, T. Jeffrey (2015): Accounting consistency and earnings quality, in: The Accounting Review.

Price, S. Mac Kay/Doran, James S./Peterson, David R./Bliss, Barbara A. (2012): Earnings conference calls and stock returns: The incremental informativeness of textual tone, in: Journal of Banking and Finance, Vol. 36, pp. 992-1011.

Purda, Lynnette/Skillicorn, David (2015): Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection, in Contemporary Accounting Research, Vol. 32, pp. 1193-1223.

Rennekamp, Kristina (2012): Processing Fluency and Investors´ Reactions to Disclosure Readability, in: Journal of Accounting Research, Vol. 50, pp. 1319-1354.

References (cont'd)

Rogers, Jonathan L./Van Buskirk, Andrew/Zechman, Sarah L. C. (2011): Disclosure Tone and Shareholder Litigation, in: The Accounting Review, Vol. 86, pp. 2155-2183.

SEC (1998): Securities and Exchange Commission, 1998b. A plain English handbook: How to create clear SEC disclosure. SEC Office of Investor Education and Assistance. http://www.sec.gov/pdf/handbook.pdf.

Smith, Malcolm/Taffler, Richard (1992): Readability and Understandability - Different Measures of the Textual Complexity of Accounting Narrative, in: Accounting, Auditing & Accountability Journal, Vol. 5, pp. 84-100.

Solomon, David H./Soltes, Eugene/Sosyura, Denis (2014): Winners in the spotlight: Media coverage of fund holdings as a driver of flows, in: Journal of Financial Economics, Vol. 113, pp. 53-72.

Sydserff, Robin/Weetman, Pauline (1999): A texture index for evaluating accounting narratives – An alternative to readability formulas, in: Accounting, Auditing & Accountability Journal, Vol. 12, pp. 459-488.

Tetlock, Paul C. (2007): Giving Content to Investor Sentiment: The Role of Media in the Stock Market, in: Journal of Finance, Vol. LXII, pp. 1139-1168.

Tetlock, Paul C./Saar-Tsechansky, Maytal/Macskassy, Sofus (2008): More Than Words: Quantifying Language to Measure Firms´ Fundamentals, in: Journal of Finance, Vol. LXIII, pp. 1437-1467.

Vrij, Aldert (2008): Detecting Lies and Deceit: Pitfalls and Opportunities.

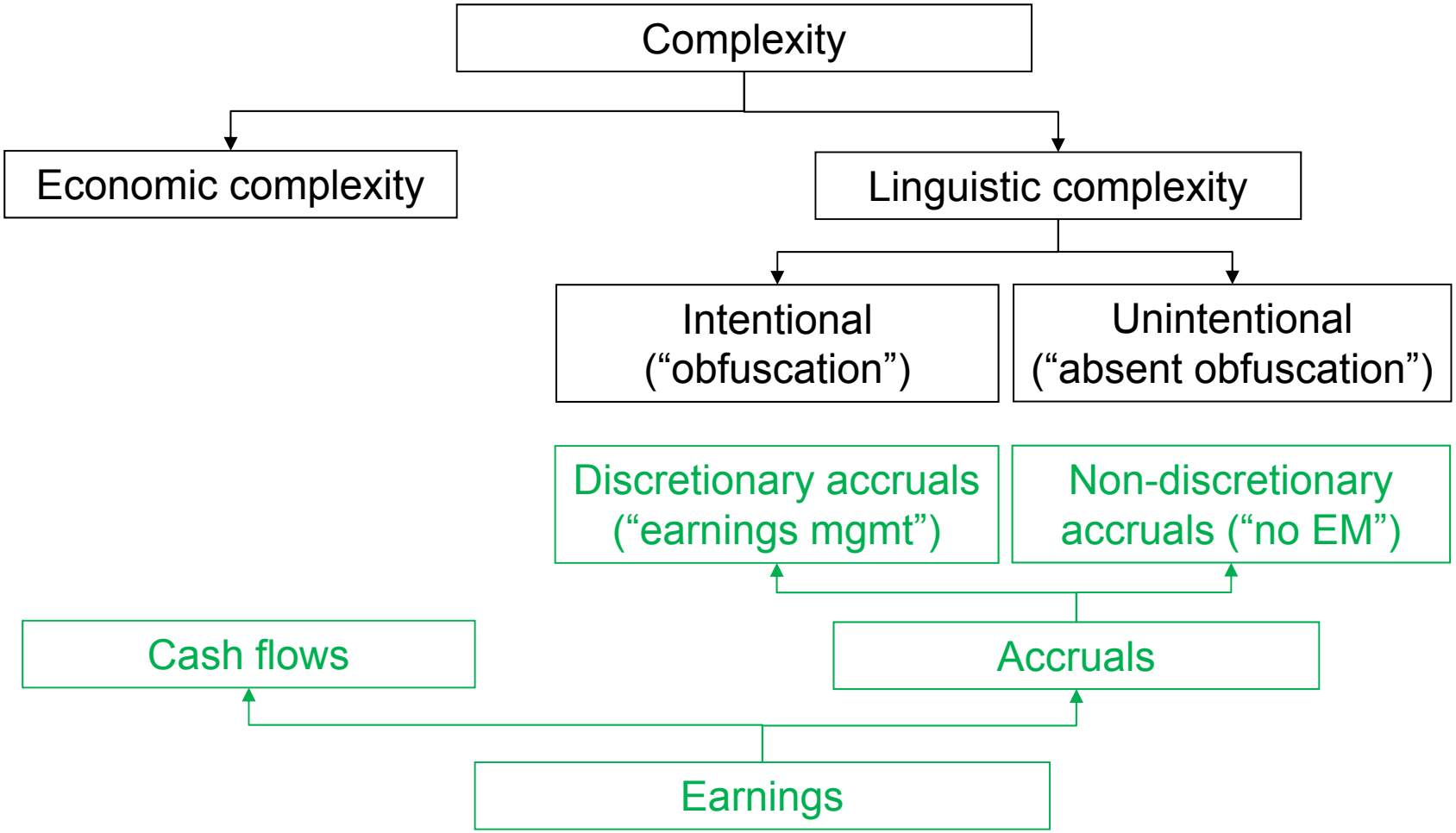
References (cont'd)

Weiss Hanley, Kathleen/Hoberg, Gerard (2010): The Information of IPO Prospectuses, in: The Review of Financial Studies, Vol. 23, pp. 2821-2864.

You, Haifeng/Zhang, Xiao-jun (2009): Financial reporting complexity and investor underreaction to 10-K information, in: Review of Accounting Studies, Vol. 14, pp. 559-586.

Backup

Distinguishing economic complexity and linguistic complexity [1)]



1) See, for example, Bushee Gow Taylor (2015 *SSRN*); Loughran McDonald (2014 *JF*); Bloomfield (2008 *JAE*; discussion of Li 2008 *JAE*)

- **Concrete vs abstract language** [1)]
  - high-level conceptual (i.e., abstract) vs lower-level, more detailed, quantified (i.e., concrete) language
  - Holding readability constant (Plain English)

ABSTRACT HIGHLIGHTED

OPERATIONS

A **We are the largest provider of digital map data for in-dash automotive navigation systems in nineteen countries worldwide.**

- **We provide wireless location-based solutions via strategic partnerships with leading content providers and mobile operators.**
- **We offer a wide range of customized public sector and enterprise solutions.**

We operate in the digital mapping, navigation and location-based services sectors. C We hold 48.7% of the in-dash automotive navigation systems market in China, 63.1% in South Africa, and 55.9% in Brazil, and the largest share of the market in 16 other countries. Collaborations with content providers, including Seeker Inc., the world's largest provider of mobile applications, and wireless service providers, including China Cellular, provide location-based services to end users via more than 14,500 websites. Customized products include aerial photogrammetry solutions, 3-D modeling applications, and location-based public sector and enterprise solutions.

CONCRETE HIGHLIGHTED

OPERATIONS

C **We hold 48.7% of the in-dash automotive navigation systems market in China, 63.1% in South Africa, and 55.9% in Brazil, and the largest share of the market in 16 other countries.**

- **Collaborations with content providers, including Seeker Inc., the world's largest provider of mobile applications, and wireless service providers, including China Cellular, provide location-based services to end users via more than 14,500 websites.**
- **Customized products include aerial photogrammetry solutions, 3-D modeling applications, and location-based public sector and enterprise solutions.**

We operate in the digital mapping, navigation and location-based services sectors. A We are the largest provider of digital map data for in-dash automotive navigation systems in nineteen countries worldwide. We provide wireless location-based solutions via strategic partnerships with leading content providers and mobile operators. We offer a wide range of customized public sector and enterprise solutions.

1) Elliott Rennekamp White (2015 *RASt*)

B 1. Text attributes: Other (cont'd)

- **Vivid vs pallid language** [1)]
  - Language that is "(a) emotionally interesting, (b) concrete and imagery-provoking, and (c) proximate in a sensory, temporal, or spatial way" as opposed to being more bland, sterile, or less emotionally charged"

- **CEO integrity** [2)]
  - Idea: Low-integrity CEOs use a dilution strategy when communicating, which manifests empirically as relatively excessive explanations (causation words) in the annual shareholder letter

- **Ethics-related terms** [3)]
  - Firms using ethics-related terms are more likely to be ''sin'' stocks, are more likely to be the object of class action lawsuits, and are more likely to score poorly on measures of corporate governance.

1) Hales Kuang Venkataraman (2011 *JAR*) | **2)** Dikolli Keusch Mayew Steffen (2014 *SSRN*) | **3)** Loughran McDonald Yun (2009 *JBusEth*)

Simple example of document classification

- Simple example, separate blog posts about earnings news (SeekingAlpha and MoneyBeat) from non-business news (ET-Online, ESPN, Politico blogs)
- Compute unadjusted term frequencies
- Use term-frequencies as inputs to a Naïve Bayes Classifier

$$P(y|x_1, \dots . x_n) = \frac{P(y)(P(x_1, \dots, x_n\ |y)}{(P(x_1, \dots, x_n\ )}$$

- y is 1 if business news and 0 if other. x_n is the occurrence of a term

Under naive assumptions classification simplified to:

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, P(y) \prod^{n} P(x_i|y)$$

- Used before in accounting research (e.g., Li 2010)

Results (Accuracy = 0.96)

Top most discerning terms for business = 1				Top most discerning terms for business = 0			
	prob(x_n|$y = 0$)	prob(x_n|$y = 1$)	prob ratio		prob(x_n|$y = 0$)	prob(x_n|$y = 1$)	prob ratio
stock	0.020%	1.097%	54.92	ap photo	0.116%	0.003%	0.027
investor	0.018%	0.969%	52.62	pm	0.291%	0.013%	0.044
compani	0.081%	2.053%	25.42	nba	0.189%	0.006%	0.034
growth	0.031%	1.134%	36.89	quarterback	0.124%	0.003%	0.025
price	0.043%	1.320%	30.71	sen.	0.148%	0.004%	0.028
invest	0.032%	0.910%	28.67	obama	0.276%	0.011%	0.039
quarter	0.040%	0.998%	25.12	photo	0.205%	0.006%	0.031
dividend	0.011%	0.407%	38.32	nfl	0.207%	0.006%	0.027
portfolio	0.010%	0.349%	35.75	, april	0.450%	0.019%	0.043
revenu	0.017%	0.453%	27.43	hi	2.278%	0.229%	0.101
custom	0.018%	0.446%	25.53	senat	0.337%	0.008%	0.024
profit	0.012%	0.342%	29.25	playoff	0.213%	0.004%	0.017
product	0.058%	0.928%	16.02	republican	0.325%	0.007%	0.022
asset	0.013%	0.367%	27.71	coach	0.308%	0.006%	0.020
compani 's	0.012%	0.330%	27.65	clinton	0.345%	0.007%	0.021
oil	0.026%	0.502%	19.49	©	0.258%	0.004%	0.015
sale	0.028%	0.520%	18.80	politico llc	0.258%	0.003%	0.012
rate	0.067%	0.911%	13.70	© politico	0.258%	0.003%	0.012
transcript	0.011%	0.263%	24.20	pm pdt	0.289%	0.003%	0.011
higher	0.024%	0.435%	18.24	politico	0.327%	0.003%	0.010

Example: Finding organizations in a document

Rule: TheGazOrganization
Priority: 50
// Matches "The <in list of company names>"
({Part of speech = DT | Part of speech = RB} {DictionaryLookup = organization})
→ Organization

Rule: LocOrganization
Priority: 50
// Matches "London Police"
({DictionaryLookup = location | DictionaryLookup = country} {DictionaryLookup = organization} {DictionaryLookup = organization}?) → Organization

Rule: INOrgXandY
Priority: 200
// Matches "in Bradford & Bingley", or "in Bradford & Bingley Ltd"
({Token string = "in"})
({Part of speech = NNP}+ {Token string = "&"} {Orthography type = upperInitial}+ {DictionaryLookup = organization end}?):orgName → Organization=:orgName

Rule: OrgDept
Priority: 25
// Matches "Department of Pure Mathematics and Physics"
({Token.string = "Department"} {Token.string = "of"} {Orthography type = upperInitial}+ ({Token.string = "and"} {Orthography type = upperInitial}+)?) → Organization

FIGURE 2. EXAMPLE RULES IDENTIFYING COMPANY NAMES IN GATE

Source: Sarawagi [2008] p.286: "Fig2.1. A Subset of rules for identifying company names paraphrased from the Named Entity recognizer in Gate."

Text attribute 1: Sentiment (cont'd)

Applied in: Giving Content to Investor Sentiment: The Role of Media in the Stock Market – Paul Tetlock (JF, 2007)

- Analyzes sentiment of the WSJ "Abreast of the market column" and relates it to market returns
- '"Abreast of the Market" column reads like a post-mortem of the market's life on the prior day' (Tetlock 2007, p. 1147)
- Finding: high media pessimism associated with downward pressure on market prices next day, followed by a reversion within the next week
- Performs a principal components factor analysis of 77 Harvard GI categories.
  - Collapses the 77 categories into a single media factor that captures the maximum variance in the GI categories (about as much as 6.5 other factors).
  - Negative and Weak GI categories each can explain over 57% of the variance in this single media factor – names it the pessimism factor
  - Finds that negative words have a much stronger correlation with stock returns than other words