TRIPLE-ACCREDITED, WORLD-RANKED

Lancaster University Management School

Quantifying accounting disclosures with textual analysis

Session 1:

Initial steps in textual analysis

Steven Young

(Lancaster University)

6th WHU Doctoral Summer Program in Accounting Research

Current Issues in Empirical Financial Reporting Research

11-14 July, 2016

SESSION 1: INITIAL STEPS

Session objectives

- Introduce the concept of textual analysis and explain why it's potentially important in the context of accounting and financial markets

- Review (some of) the main textual analysis methods applied to date in the accounting literature

- Emphasize the limitations and dangers with applying automated textual analysis to study qualitative disclosures in accounting and finance

SESSION 1: INITIAL STEPS

Context

- Analysis of qualitative information has a long tradition in computer science (natural language processing) and linguistics (corpus linguistics)
- Analysis of language (spoken and written) can provide powerful insights:
  - Digital identities
  - Dementia screening
  - Alternative approach to studying economic consequences
- Methods only recently started to gain traction in accounting and finance
  - Earlier work on disclosures involved manual analysis of small samples → concerns over objectivity and generalizability at top US journals
- Not before time…
  - Estimates suggest 90% of all available data created in last 10 years, 80% of which in a business context is qualitative/unstructured
  - Rapid growth in nontraditional information sources (Tweets, blogs, etc.)

SESSION 1: INITIAL STEPS

Textual analysis

Harvest text

Extract & clean text

Analyse text

More difficult when content is unstructured and provided in inaccessible file formats (e.g., annual reports as PDF files)

Relatively straightforward when content follows consistent structure and is presented in ASCII format (e.g., 10-K filings via EDGAR)

- Word-level analysis (dictionaries, collocates)
- POS & semantic tagging to capture meaning
- Construct & analyze corpora (identify key words & themes)
- Text classification (statistical model to categorize content)
- Text mining to identify patterns (AI methods)

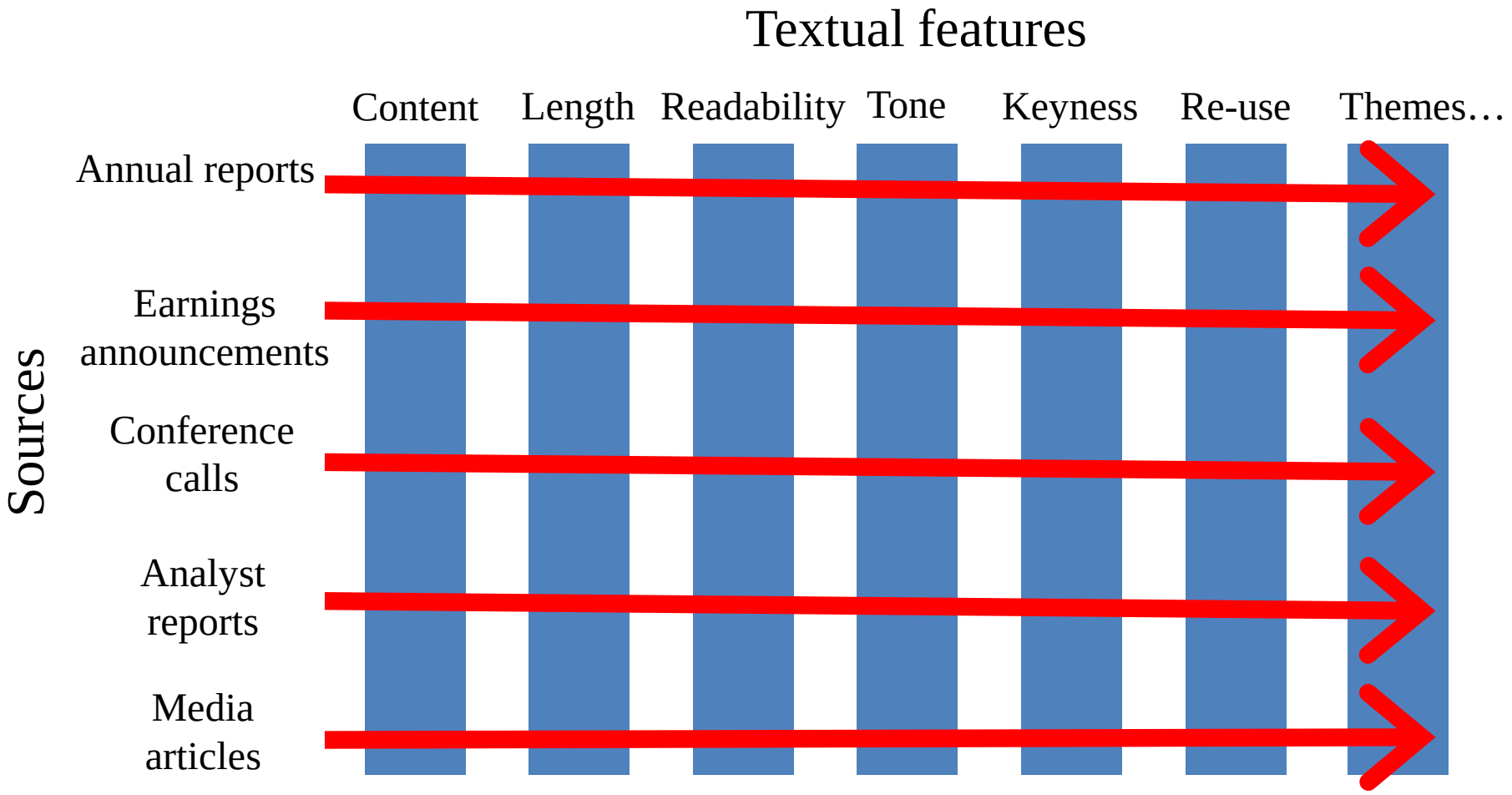
SESSION 1: INITIAL STEPS

Harvesting, extracting & cleaning text

- Application of NLP and corpus methods involves large volumes of text
  - SEC filings via EDGAR
  - Annual reports via *Thomson*, *Perfect Information*, directly from websites…
  - Media articles via *Factiva* or directly via publication API
  - Analyst reports and conference call transcripts via *Thomson StreetEvents*
  - Tweets (e.g., via Twitter), message posting (e.g., via Seeking Alpha), blogs…
- Accounting researchers often use pearl script to harvest and process documents
  - Andrew Leone provides pearl resources for accessing EDGAR
  - Other languages as good or better (e.g., python, java, etc.)
  - Area where comparative advantage is important
- Some types of qualitative data are easier to access than others → risk that attention focuses on what's accessible

SESSION 1: INITIAL STEPS

Analyzing text



SESSION 1: INITIAL STEPS

Analyzing text: Overview

- Methods applied in extant accounting research tend to operate at the **individual word-level**
  - Unit of analysis is an individual word rather than a group of words (e.g., statement, sentence or paragraph)

> *The combination of a long-term decline in drinking-out of approximately 3.5% per annum, changing customer behaviour, relative price positioning and the impact of regulation means that the number of pubs in the UK is expected to continue to decline*

  - How positive or optimistic is the statement?
  - Is the statement focused on the present, past, or future?
  - Is the statement easy to understand?

SESSION 1: INITIAL STEPS

Analyzing text: Overview *cont.*

- Methods applied in extant accounting research tend to operate at the **individual word-level**
  - Unit of analysis is an individual word rather than a group of words (e.g., statement, sentence or paragraph)

- Word-level approaches appearing in the accounting literature include:
  - Dictionary methods → count the number of words from a pre-defined dictionary that captures a specific aspect (e.g., positivity, forward-looking)
  - Readability and complexity methods → count the number of words in a sentence or the entire document; count the number of complex words
  - Text similarity → proportion of words in a statement by Firm A at time *t* also appear in the corresponding statement by Firm B (or by Firm A in *t* + 1)?

- Termed **bag-of-words** approaches because words considered in isolation from their context, meaning, grammatical usage, etc.

SESSION 1: INITIAL STEPS

Bag-of-words methods: Dictionaries

- Based on wordlists (dictionaries) designed to capture any specific construct (e.g., positivity, negativity, uncertainty, forward-lookingness, …)

- Approach 1 → Comprehensive dictionaries capturing as many different words as possible relating to a particular theme (e.g., positivity)
  - ✓ Objective and replicable (e.g., wordlists in General Inquirer, Diction, etc.)
  - × Lack information on context and meaning → no **disambiguation**

Example: "bank" has multiple uses and meanings
1. Financial institution *(noun)*
2. Element of currency (*noun*)
3. Part of a river *(noun)*
4. Public holiday in the UK *(noun)*
5. To deposit *(verb)*
6. To rely on *(verb)* …

SESSION 1: INITIAL STEPS

Bag-of-words methods: Dictionaries

- Based on wordlists (dictionaries) designed to capture any specific construct (e.g., positivity, negativity, uncertainty, forward-lookingness, …)

- Approach 1 → Comprehensive dictionaries capturing as many different words as possible relating to a particular theme (e.g., positivity)
  - ✓ Objective and replicable (e.g., wordlists in General Inquirer, Diction, etc.)
  - × Lack information on context and meaning → no disambiguation

Example: further complicated when inflections are considered
1. banking
2. banker
3. banked
4. bankable…

SESSION 1: INITIAL STEPS

Bag-of-words methods: Dictionaries

- Based on wordlists (dictionaries) designed to capture any specific construct (e.g., positivity, negativity, uncertainty, forward-lookingness, …)

- Approach 1 → Comprehensive dictionaries capturing as many different words as possible relating to a particular theme (e.g., positivity)
  - ✓ Objective and replicable (e.g., wordlists in General Inquirer, Diction, etc.)
  - × Lack information on context and meaning → no disambiguation

- Approach 2 → Contextual dictionaries developed by the researcher for use in a specific setting
  - Disambiguation → attempt to recognize that the same word may have different meanings in different settings
  - See Loughran & McDonald (2011) → presentation
  - Only deals partially with the problem of context and meaning → e.g., ignores **part of speech** (noun, verb, preposition, conjunction…)

SESSION 1: INITIAL STEPS

Bag-of-words methods: Readability and complexity

- Ease of understanding for English writing
  - Based on view that using more words and longer words makes the text more difficult to understand, *all else equal*
  - Often used in the accounting literature as a proxy for obfuscation (Li 2008)

- Approach 1 → direct attempt to measure sentence complexity such as **Fog Index** (Günning 1968)

$$Readability = 0.4\left(\left[\frac{words}{sentences}\right] + 100 \times \left[\frac{complex\ words}{words}\right]\right)$$

complex words are words of ≥ 3 syllables, excl. proper nouns, jargon, common suffixes (-es, -ed, -ing)

  - Not all complex words are difficult → *internationalization* has 8 syllables
  - Short sentences containing short words is no guarantee reading is easier
  - Index applies to general writing rather than financial text
  - Algorithmic approach fails to reflect differences in actual meaning

SESSION 1: INITIAL STEPS

Bag-of-words methods: Readability and complexity

Example 1: *Microsoft delivered lower-than-expected sales revenues and lower profits due to supply-chain problems*

Example 2: *Microsoft delivered lower-than-expected sales revenues (due to supply-chain problems) and lower profits*

- Two examples have identical Fog scores (because words are identical) but **Example 1 is more ambiguous** and hence harder to understand
  - Fog algorithm treats text as a bag-of-words

SESSION 1: INITIAL STEPS

Bag-of-words methods: Readability and complexity *cont.*

- Approach 2 → indirect proxies for complexity and ease of reading
  - Document length → number of words or number of pages (Li 2008)
  - File size → 10-K filings (Loughran & McDonald 2014)

- Based on the assumption that longer communications are more difficult to read and understand
  - Correlation between 10-K file size and Fog index = 0.37

- Approach takes no account of textual content
  - Unlikely to be viewed by linguists and computer scientists as being a reliable means of measuring the complexity and understandability of annual reports

SESSION 1: INITIAL STEPS

Bag-of-words methods: Text similarity

- Similarity of language between two or more sections of text
  - Cross-sectional comparisons → compare section *i* for Firms A and B in year *t*
  - Time-series comparisons → compare section *i* for Firm A in years *t* and *t* + 1 (Brown & Tucker 2011)

- Approach → Cosine similarity aims to provide a representation of the text that takes account of all information contained therein
  - For example, count # of times a word type appears in text → $vector_{A\,t}$

SESSION 1: INITIAL STEPS

Bag-of-words methods: Text similarity

The combination of a long-term decline in drinking-out of approximately 3.5% per annum, changing customer behaviour, relative price positioning and the impact of regulation means that the number of pubs in the UK is expected to continue to decline

	$vector_t$
of	4
the	4
decline	2
in	2
to	2
a	1
and	1
annum	1
approximately	1
behaviour	1
changing	1
combination	1
continue	1
customer	1
drinking-out	1
expected	1
impact	1
is	1
long-term	1
means	1
number	1
per	1
positioning	1
price	1
pubs	1
regulation	1
relative	1
that	1
UK	1

SESSION 1: INITIAL STEPS

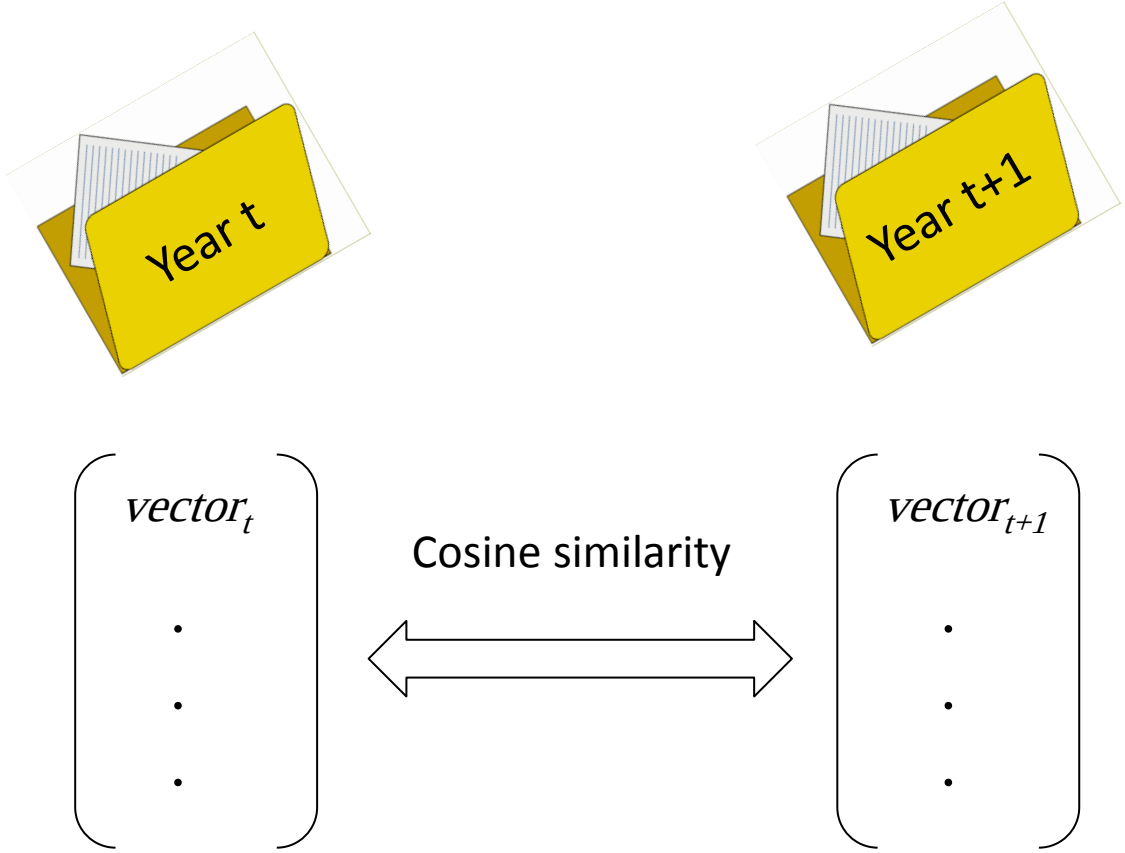
Bag-of-words methods: Text similarity

- Similarity of language between two or more sections of text
  - Cross-sectional comparisons → compare section *i* for Firms A and B in year *t*
  - Time-series comparisons → compare section *i* for Firm A in years *t* and *t* + 1 (Brown & Tucker 2011)

- Approach → Cosine similarity aims to provide a representation of the text that takes account of all information contained therein
  - For example, count # of times a word type appears in text → $vector_{A\ t}$
  - Repeat for second document to generate comparable vector → $vector_{A\ t+1}$
  - Compare similarity of two or more vectors by computing cosine similarity:

SESSION 1: INITIAL STEPS

Bag-of-words methods: Text similarity



SESSION 1: INITIAL STEPS

Bag-of-words methods: Text similarity

- Similarity of language between two or more sections of text
  - Cross-sectional comparisons → compare section *i* for Firms A and B in year *t*
  - Time-series comparisons → compare section *i* for Firm A in years *t* and *t* + 1 (Brown & Tucker 2011)

- Approach → Cosine similarity aims to provide a representation of the text that takes account of all information contained therein
  - For example, count # of times a word type appears in text → $vector_{A\,t}$
  - Repeat for second document to generate comparable vector → $vector_{A\,t+1}$
  - Compare similarity of two or more vectors by computing cosine similarity:

$$Cos\,\theta = \frac{v_{A\,t} \cdot v_{A\,t+1}}{\| v_{A\,t} \| \| v_{A\,t+1} \|}$$

  - Cos 90 = 0 ⇒ vectors not similar
  - Higher cosine value ⇒ higher similarity (in terms of word frequencies)

SESSION 1: INITIAL STEPS

Analyzing text: Statistical methods

- Significant element of computational linguistics/NLP involves building statistical models to:
  - Identify interesting patterns in unstructured text → e.g., daily measure of national happiness based on aggregate Facebook posts
  - Discover new knowledge from these patterns → e.g., the type of language that predicts financial fraud
  - Automatically classify text into distinct categories → e.g., positive statements vs. negative statements

- Statistical approaches in NLP include:
  - Text classifiers (machine learning) (Li 2010, Huang et al. 2014, Lee et al. 2014)
  - Text mining (Balakrishnan et al. 2010, Chen et al. 2013)
  - Information extraction (Zaki & Theodoulidis 2013)

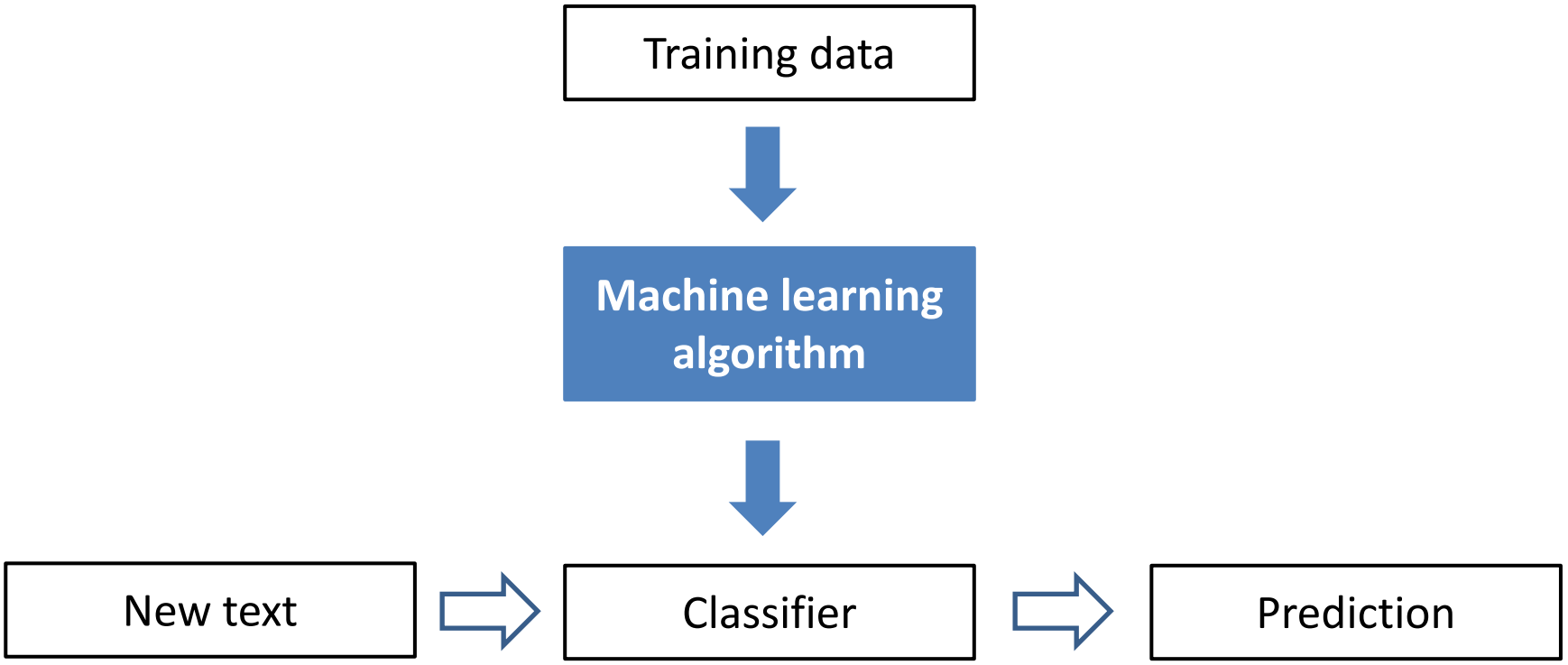
SESSION 1: INITIAL STEPS

Text classification/categorization

- Automatically classify any given text (sentence, section, entire document) into a predefined class set comprising two or more elements
  - Avoids reading and scoring text manually, thereby facilitating consistent, objective large sample analysis
  - Common example is naïve Bayes classifier (Li 2010, Huang et al. 2014)

SESSION 1: INITIAL STEPS

Text classification/categorization



Source: Raschka (2014 http://arxiv.org/pdf/1410.5329.pdf)

SESSION 1: INITIAL STEPS

Text classification/categorization

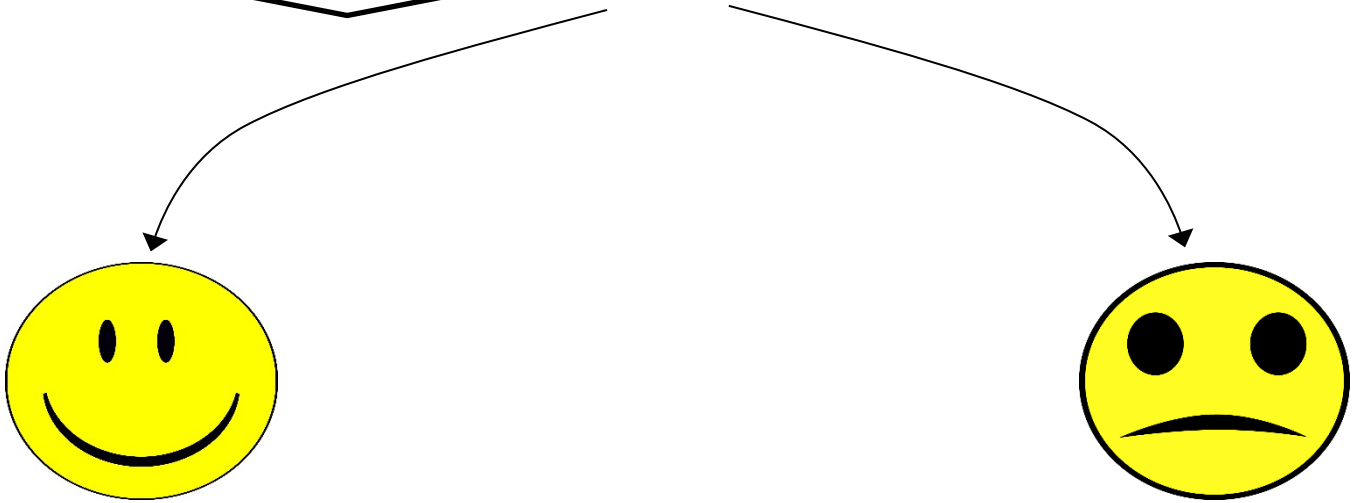
- Automatically classify any given text (sentence, section, entire document) into a predefined class set comprising two or more elements
  - Avoids reading and scoring text manually, thereby facilitating consistent, objective large sample analysis
  - Common example is naïve Bayes classifier (Li 2010, Huang et al. 2014)

- Approach → naïve Bayes provides a model for classifying text into groups based on conditional probabilities
  - Define a "training dataset" where manual coders link text feature ***x*** (e.g., words from a dictionary) with outcome category ***y*** (e.g., positivity, negativity)

SESSION 1: INITIAL STEPS

Example of text classification/categorization

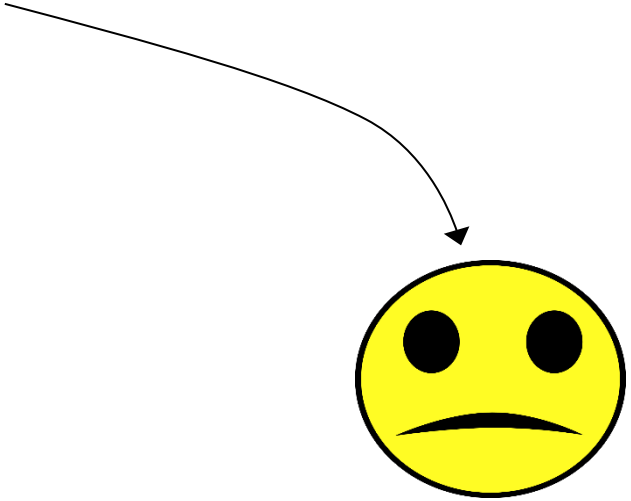
The combination of a long-term decline in drinking-out of approximately 3.5% per annum, changing customer behaviour, relative price positioning and the impact of regulation means that the number of pubs in the UK is expected to continue to decline



SESSION 1: INITIAL STEPS

Example of text classification/categorization

The combination of a long-term decline in drinking-out of approximately 3.5% per annum, changing customer behaviour, relative price positioning and the impact of regulation means that the number of pubs in the UK is expected to continue to decline



SESSION 1: INITIAL STEPS

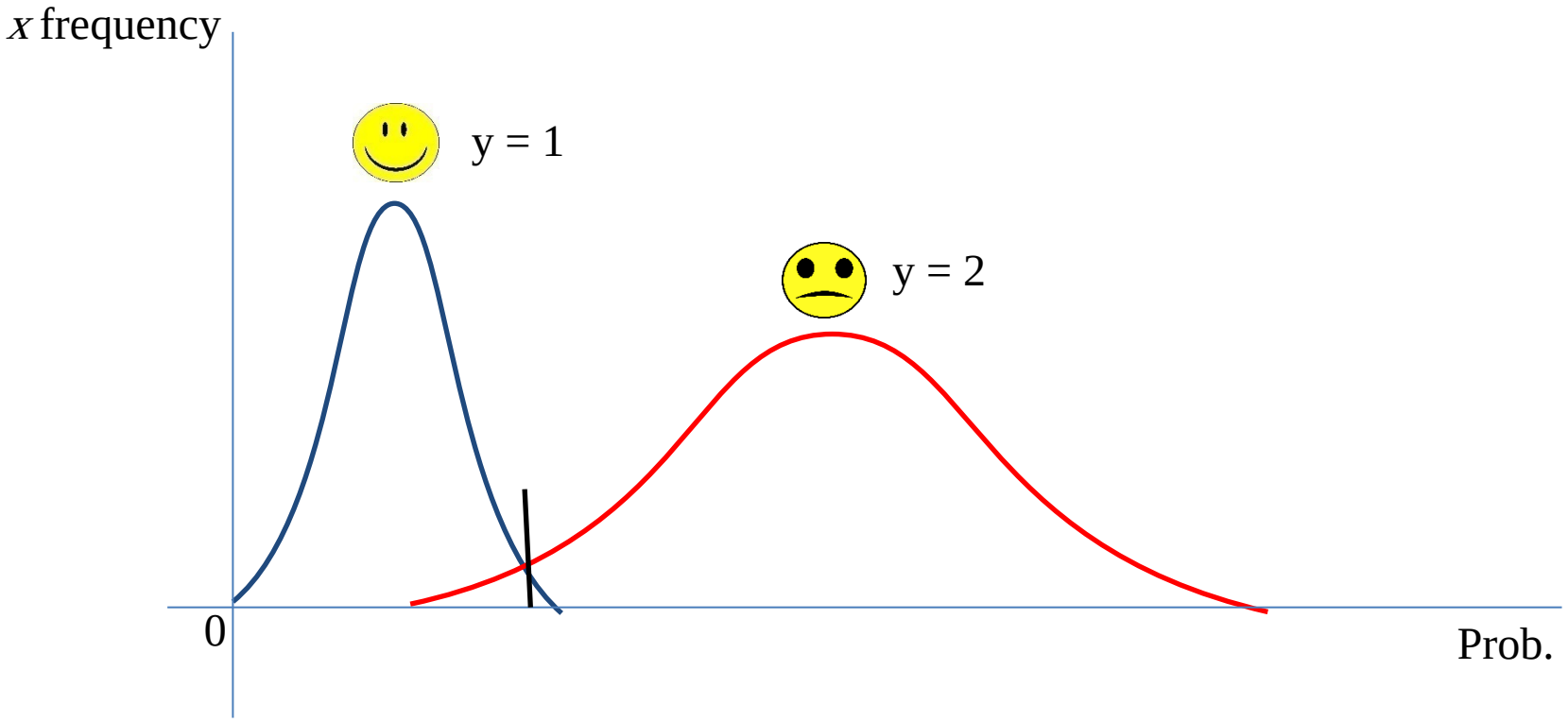
Text classification/categorization

- Automatically classify any given text (sentence, section, entire document) into a predefined class set comprising two or more elements
  - Avoids reading and scoring text manually, thereby facilitating consistent, objective large sample analysis
  - Common example is naïve Bayes classifier (Li 2010, Huang et al. 2014)

- Approach → naïve Bayes provides a model for classifying text into groups based on conditional probabilities
  - Define a "training dataset" where manual coders link text feature ***x*** (e.g., words from a dictionary) with outcome category ***y*** (e.g., positivity, negativity)
  - Model joint probability p(*x*,*y*) using observations of *x* and *y* from training data → compute class conditional densities

SESSION 1: INITIAL STEPS

Class conditional densities based on training data



SESSION 1: INITIAL STEPS

Text classification/categorization

- Automatically classify any given text (sentence, section, entire document) into a predefined class set comprising two or more elements
  - Avoids reading and scoring text manually, thereby facilitating consistent, objective large sample analysis
  - Common example is naïve Bayes classifier (Li 2010, Huang et al. 2014)

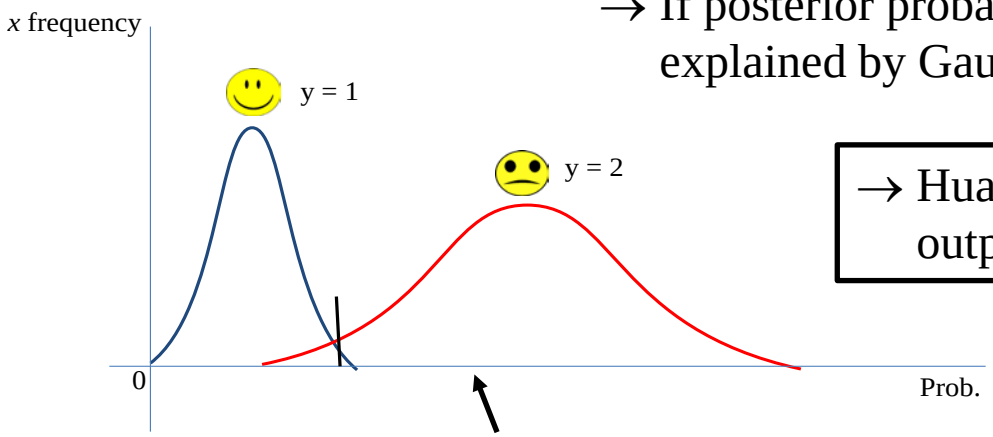
- Approach → naïve Bayes provides a model for classifying text into groups based on conditional probabilities
  - Define a "training dataset" where manual coders link text feature ***x*** (e.g., words from a dictionary) with outcome category ***y*** (e.g., positivity, negativity)
  - Model joint probability p(x,y) using observations of x and y from training data → compute class conditional densities
  - Use Bayes rule to compute p(y|x) for unread text where x is extracted automatically → compare with observed class conditional densities

SESSION 1: INITIAL STEPS

Example of text classification/categorization

Sales and profits in the period where down against expectations following disappointing Christmas trading and increased costs due to unfavourable exchange rate movements

- Apply Bayes rule to compute posterior probability $p(y = 2 | x = 4)$



→ If posterior probability for $y = 2$ is high then unread text better explained by Gaussian distribution for $y = 2$ (negativity)

→ Huang et al. (2014) show naïve Bayes outperforms dictionaries for sentiment measure

SESSION 1: INITIAL STEPS

Note of caution

- Increasing ease with which unstructured data can be analyzed creates **threats** as well as opportunities
- Risk of analyzing what's easy or available rather than what's interesting or economically important
  - Just because you can doesn't mean you should!
  - Research idea, theory and incremental contribution are always the primary determinants of success

SESSION 1: INITIAL STEPS

Note of caution

- Increasing ease with which unstructured data can be analyzed creates **threats** as well as opportunities
- Risk of analyzing what's easy or available rather than what's interesting or economically important
  - Just because you can doesn't mean you should!
  - Research idea, theory and incremental contribution are always the primary determinants of success

- Don't get seduced by quasi-rigor and apparent application of the "scientific method"
  - The ability to process thousands of documents does not guarantee the research is either **relevant** or **reliable**
- Understand your comparative advantage

TRIPLE-ACCREDITED, WORLD-RANKED

Lancaster University Management School

Quantifying accounting disclosures with textual analysis

Session 2:

Next steps in textual analysis

Steven Young

(Lancaster University)

6th WHU Doctoral Summer Program in Accounting Research

Current Issues in Empirical Financial Reporting Research

11-14 July, 2016

SESSION 2: NEXT STEPS

Session objectives

- Increase awareness of mainstream NLP and corpus methods that are yet to feature in the accounting literature
- Highlight potential research opportunities resulting from improved information retrieval
- Stress how automated textual analysis methods are unlikely to replace manual coding in all areas of accounting research

Session 2: Next Steps

Context

- Literature in accounting and finance has only scratched the surface of textual analysis capabilities
  - Reliance on basic NLP techniques primarily involving bag-of-words methods
  - Little use of corpus methods
- > 20-30 years behind developments in computational linguistics and machine learning
- Lagging behind other business disciplines where application of computational linguistics approaches has a longer tradition
  - Strategy → *Strategic Management Journal*
  - Management → *Academy of Management; Administrative Science Quarterly*
  - Marketing → *Journal of Marketing; Journal of Marketing Research*
  - Management Science/OR → *Management Science; European Journal of Operational Research*

SESSION 2: NEXT STEPS

NLP topics

Machine Translation and Evaluation
Sentiment Analysis and Emotion Recognition
Corpora for Language Analysis
Information Extraction and Retrieval
Multimodality
Multiword Expressions
Named Entity Recognition
Parsing
Summarisation
Word Sense Disambiguation
Multilingual Corpora
Lexicons
Semantics
Sentiment Analysis and Opinion Mining
Treebanks

Document Classification & Text Categorisation
Morphology
Multimodality
Ontologies
Part of Speech Tagging
Tweet Corpora and Analysis
Twitter-Related Analysis
Social Media
Word Sense Disambiguation
Prosody and Phonology
Crowdsourcing
Corpus Querying and Crawling
Grammar and Syntax
Parallel and Comparable Corpora

SESSION 2: NEXT STEPS

Refined NLP methods

- Dictionaries → Greater emphasis on domain-specific wordlists following Loughran & McDonald (2011)
  - Tone (Henry & Leone 2016)
  - Causation (Dikolli et al. 2014)
  - Strategy (Anathaskou et al. 2016)

- Readability → Develop accounting-specific measures
  - Notions of readability and understandability in a financial reporting context are likely to differ from less technical contexts where Fog Index developed

- Word-level analysis → Part-of-speech (POS) tagging
  - POS tagging designed to differentiate between nouns, verbs, adjectives, prepositions, conjunctions, etc. to improve meaning and context
  - Allocation to POS category based on definition and context (i.e., relationship with adjacent and related words)

SESSION 2: NEXT STEPS

Disambiguation with POS tagging

- Example: "power" is included in Loughran & McDonald's *positive* wordlist

ITM Power are now in a position where we can focus on delivering its leading refuelling and energy storage products

R&D is critical to achieving market power in an increasing competitive marketplace

Our strategy is enabling the business to power ahead relative to the competition

- Free POS tagging toolkits are available including:
  - Stanford Log-linear POS Tagger: http://nlp.stanford.edu/software/tagger.shtml
  - CLAWS POS tagger http://ucrel.lancs.ac.uk/claws/

SESSION 2: NEXT STEPS

Refined NLP methods *cont.*

- Semantic analysis → the structure and meaning of text
  - Identifying grammatical patterns to uncover meaning from the author's perspective
  - More advance tool for disambiguation

- Example: Google search for "*jaguar*" returns multiple options:



  - Semantic analysis looks for words and phrases that distinguish webpages about cars from those about big cats, American football, font type, etc.
  - Semantic taggers are available → http://ucrel.lancs.ac.uk/usas/

SESSION 2: NEXT STEPS

Refined NLP methods *cont.*

- Machine learning applications for classification are limited in accounting and finance (Li 2010)

- Exclusive focus on supervised generative classification via naïve Bayes:
  - ✓ Advantages: easy to train and understand results; many different extensions exist; fast; good performance on average
  - × Weaknesses: naïve assumptions that may not reflect the data; risk of overfitting

- Alternative classification methods exist:
  - Generative → **Random Forest** is potentially more robust to overfitting than naïve Bayes but harder to train
  - Discriminative → **Fisher's Linear Discriminant, Logistic Regression** and **support vector machines (SVM)** seek to model $p(y|x)$ directly
  - Unsupervised → classify text with no prior training

SESSION 2: NEXT STEPS

Refined NLP methods *cont.*

- Cosine similarity is a simplistic tool for comparing text → content can be identical even though individual words differ

> *If markets react less completely to information that is less easily extracted from public disclosures, then managers have more incentive to obfuscate information when firm performance is bad* (Li 2008)

> *Management face incentives to engage in impression management behaviour in the wake of poor results if investors fail to respond fully to corporate communications that are hard to read and understand* (Young 2016)

- More sophisticated approaches to assessing similarity:
  - **Paraphrase identification** → recognizing text fragments with similar meaning
  - **Sematic textual similarity** → degree of semantic equivalence between texts

SESSION 2: NEXT STEPS

Refined NLP methods *cont.*

- Topic identification → **topic models** seek to uncover abstract topics or themes in a body of text
  - **Latent Dirichlet Allocation** (LDA) and **Latent Sematic Analysis** (LSA)
  - LDA used to identify key topic(s) discussed by management in MD&A (Ball et al. 2014, Dyer et al. 2016)

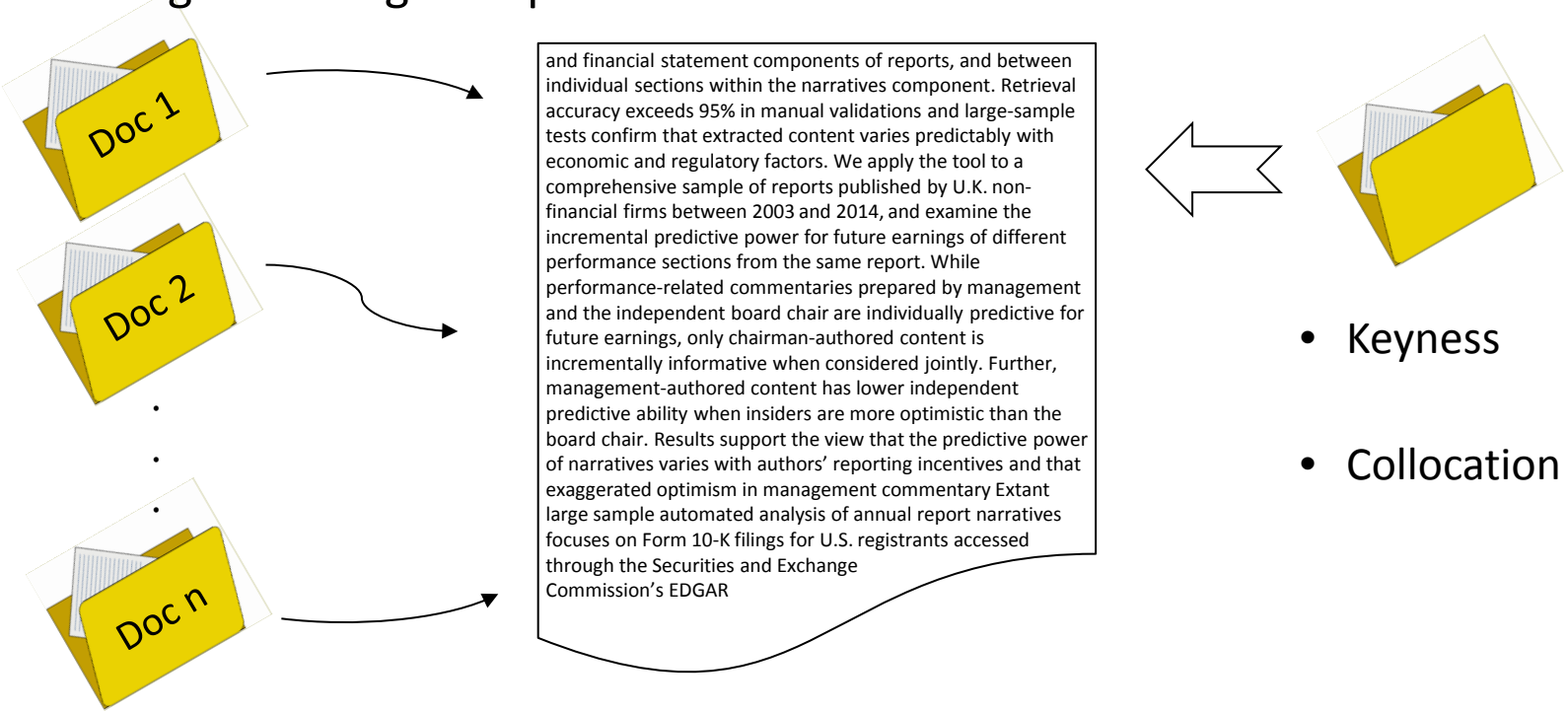
- Opinion and text mining → data-driven text classification models based on statistical relations
  - Draw on neural networks and artificial intelligence (AI)
  - No attempt to understand properties of the text → classifier is pure black-box
  - Applications in accounting and finance include Balakrishnan et al. (2010) and Chen et al. (2013)

SESSION 2: NEXT STEPS

Corpus approaches

- **Corpus linguistics** studies the properties of language as expressed in *corpora* (i.e., samples) of actual text

- Building and using a corpus:



SESSION 2: NEXT STEPS

Document processing and retrieval

- Majority of extant work focused on documents that are *relatively* straightforward to process due to format and structure
  - Form 10-Ks via EDGAR → standardized reporting templates, ASCII format
  - Conference call transcripts → standardized structure, HTML tags
  - Media articles, blogs & Tweets → relatively short, simple structure, ASCII format
  - Earnings press releases (???)

  → Accurate text retrieval and classification is feasible

- Other forms of financial communication are important → more sophisticated extraction and classification procedures required
  - PDF annual reports → no standardized structure, poor accessibility, infographics
  - Web pages → no standardized structure, dynamic, embedded content, irrelevant content
  - Comment letters → different styles, various formats, irrelevant content
  - Regulatory documents → no standardized structure, PDF files

SESSION 2: NEXT STEPS

Document processing and retrieval: PDF annual reports

- 10-Ks represent only part of U.S. firms' annual report disclosures
  - Most registrants also publish a **non-standardized** "glossy" report containing graphics, photos, and supplementary narratives (e.g., letter to shareholders)
  - Lack the consistent, linear structure of the annual report on Form 10-K
  - Typically presented as PDF files

- Outside the U.S., digital PDF annual reports are the primary (and often *only*) format in which firms present their annual report and accounts
  - Management enjoys significant discretion over information disclosed, order in which information is presented, and labels used to describe specific sections

GROUP FINANCIAL REVIEW

"Our performance this year reflected both the volatile global consumer and economic environment and the actions we took to strengthen the business. Reported net sales were up with the integration of USL and organic net sales flat driven by currency related challenges in specific emerging markets and embedding our sell out discipline. Our focus on cost delivered savings and drove margin expansion, prioritising cash resulted in a marked cash flow improvement and we continued to invest for the future.

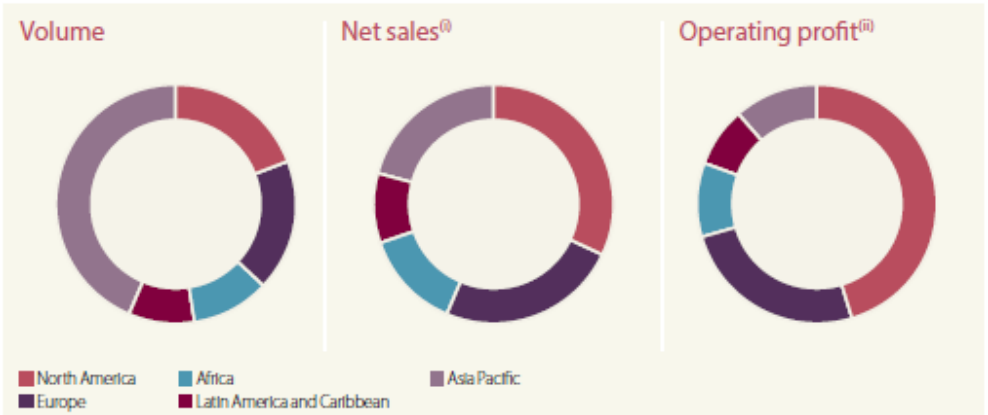
Deirdre Mahlan, Chief Financial Officer

Reported net sales up
5%
with full consolidation of United Spirits

Free cash flow of
£2bn
up £0.7 bn

9%
final dividend increase to give recommended full year dividend of 56.4 pence

Organic net sales
flat



(i) Excluding corporate net sales. (ii) Before exceptional items and corporate costs.

Key performance indicators		2015	2014
Organic net sales growth	%	–	–
Organic operating margin improvement	basis points	24	77
Earnings per share before exceptional items	pence	88.8	95.5
Free cash flow	£ million	1,963	1,235
Return on average invested capital (ROIC)(i)	%	12.3	14.1

MARKET DYNAMICS

The global beverage alcohol market is large and diverse, with an estimated six billion equivalent units of alcohol sold each year, generating £300 billion of net sales. It is also one of the most regulated in the world, and beverage alcohol companies operate in the context of a range of stakeholder expectations and demands. This environment presents opportunities for a business like Diageo, with our global scale, our diverse range of leading brands, and our high standards of governance and ethics.

A growing global market

Beverage alcohol is a profitable, growing and attractive market in which to participate. Margins are significantly higher than for the overall consumer goods market, while, over the medium term, the industry is expected to grow in both volume and value. While the global market is split almost equally between emerging and developed markets, emerging markets are expected to grow at a faster rate.

Within emerging markets and developed markets, every individual market presents different consumer dynamics and a different outlook determined by specific local conditions. Our 21-market operating model, coupled with tailored local strategies, enables us to meet the specific needs of consumers across different geographies.

Split of global total beverage alcohol (TBA) volume (EU)

Developed markets
Emerging markets

Year ended 31 December 2014
Source: Diageo estimates

Split of global total beverage alcohol (TBA) net sales (£)

Developed markets
Emerging markets

Year ended 31 December 2014
Source: Diageo estimates

buying brands and they have more money to spend on them. There is a good

hedge against individual market volatility, while tailored strategies for each market

SESSION 2: NEXT STEPS

Document processing and retrieval: PDF annual reports

- 10-Ks represent only part of U.S. firms' annual report disclosures
  - Most registrants also publish a **non-standardized** "glossy" report containing graphics, photos, and supplementary narratives (e.g., letter to shareholders)
  - Lack the consistent, linear structure of the annual report on Form 10-K
  - Typically presented as PDF files

- Outside the U.S., digital PDF annual reports are the primary (and often *only*) format in which firms present their annual report and accounts
  - Management enjoys significant discretion over information disclosed, order in which information is presented, and labels used to describe specific sections

- Unstructured and inconsistent format coupled with non-ASCII/HTML file types (e.g., PDF) creates major extraction and classification challenges
  - Lack of systematic large sample evidence despite enduring status as a key element of corporate communication

SESSION 2: NEXT STEPS

Document processing and retrieval: PDF annual reports *cont.*

- Lang & Stice-Lawrence (2015) conduct first large sample analysis of non-10-K annual reports for international sample of > 87,000 PDF reports

- Approach the problem of analysing unstructured PDF reports by:
  - Converting files to ASCII format using proprietary software
  - Isolating running text with a pearl script

- Method facilitates analysis of content at the aggregate level but fails to capture information on the location of commentary within the document
  - Unable to distinguish disclosures in the footnotes to the financial statements from commentary in the narrative component of the report
  - Unable to distinguish between disclosures from distinct sections of the narrative component
  - No information on document structure → important dimension of disclosure

SESSION 2: NEXT STEPS

Document processing and retrieval: PDF annual reports *cont.*

- Alves et al (2016) develop a software tool for extracting and classifying narrative content from digital PDF annual reports
  - Detect the page containing the annual report table of contents
  - Extract the table of contents (section titles and corresponding page numbers)
  - Synchronize page numbers in the digital PDF file
  - Use synchronized page numbers to determine start and end of each section, then extract content section by section
  - Content is partitioned into the audited financial statements component of the report and the "front-end" narratives component
  - Narratives further subclassified into generic report sections (shareholders' letter, CEO review, CFO review, governance statements, remuneration reports)
  - http://ucrel.lancs.ac.uk/cfie/wmatrix-import-software.php

SESSION 2: NEXT STEPS

Understanding the limitations

- Automated textual analysis methods provide the opportunity to develop profound new insights into financial communication

- But automated methods are unlikely to provide answers to all research questions concerning disclosure and financial communication
  - Some (many) important research questions require more refined approaches designed to detect more subtle effects

- Examples:
  - Attribution bias
  - Obfuscation and impression management

CHAIRMAN'S STATEMENT:
WE ARE CREATING ONE OF THE BEST PERFORMING, MOST TRUSTED AND RESPECTED CONSUMER PRODUCTS COMPANIES.

"Over the last two years we have taken the necessary steps to strengthen Diageo, to position our company to drive sustainable growth and value for you, our shareholders, and to ensure we are a trusted and respected partner to all our stakeholders around the world.

Interim dividend per share

21.5p (↑9%)

31 December 2013: 19.7p

Final recommended dividend per share

Diageo is a leader in beverage alcohol, one of the most attractive growth sectors in consumer products. With our portfolio of global and local brands, and the presence we have built in developed and emerging markets, we are well positioned to capture this growth.

Performance and dividend

In a volatile global environment, our total dividend for the year to 56.4 pence per share, an increase of 9% over the prior year. The final dividend will be paid to shareholders on 8 October 2015. Earnings per share to dividend cover at 1.6 times is now outside our cover ratio, and we will look to rebuild cover over time, maintaining dividend increases at a mid-single digit rate until we are back in range.

SESSION 2: NEXT STEPS

Understanding the limitations

- Automated textual analysis methods provide the opportunity to develop profound new insights into financial communication

- But automated methods are unlikely to provide answers to all research questions concerning disclosure and financial communication
  - Some (many) important research questions require more refined approaches designed to detect more subtle effects

- Examples:
  - Attribution bias
  - Obfuscation and impression management
  - Format and presentation
  - Tables, pictures and infographics
  - Quality

TRIPLE-ACCREDITED, WORLD-RANKED

Lancaster University Management School

Narrative information in financial markets

Session 3

Evidence: Annual reports and earnings announcements

Steven Young

(Lancaster University)

6th WHU Doctoral Summer Program in Accounting Research

Current Issues in Empirical Financial Reporting Research

11-14 July, 2016

SESSION 3: ANNUAL REPORTS & EARNINGS ANNOUNCEMENTS

Session objectives

- Review evidence on the properties and economic consequences of textual commentary in corporate reports and earnings announcements
- Understand when the application of large-sample textual analysis methods adds value
- Highlight the research opportunities associated with non-US disclosure environments

SESSION 3: ANNUAL REPORTS & EARNINGS ANNOUNCEMENTS

Information or obfuscation?

- Two polar views on the role of narrative reporting on which large-sample studies using textual analysis seek evidence

- Management-provided narratives are incrementally informative
  - Disclosures and commentaries complement financial statement information → provide contextual information to understand financial results
  - Disclosures and commentaries supplement financial statement information → compensate for financial reporting biases and boundaries

- Preparers use narrative disclosures to obfuscate weak financial performance → impression management
  - Opportunistic management exploit reporting discretion to present a favourable view of firm performance (broadly defined)
  - Assumes the market has limited information processing ability

SESSION 3: ANNUAL REPORTS & EARNINGS ANNOUNCEMENTS

Information

- Do narrative disclosures predict future performance?

$$Perform_{it+n} = \gamma_0 + \gamma_1 Perform_{it} + \gamma_2 Tone_{it} + \sum_{k=1}^{K} \delta_k Controls_{kit} + \varepsilon_{it+n}$$

- Consistent evidence using US data that $\hat{\gamma}_2 > 0$ when *Perform* equals earnings, ROA, cash flow, dividends, etc. using:
  - Tone of earnings press release (Davis et al. 2012, Henry & Leone 2016)
  - Tone of Chairman's letter (Abrahamson & Amir 1996, Alves et al. 2016)
  - Tone of MD&A (Li 2010, Henry & Leone 2016, Alves et al. 2016)

- Predictive ability extends beyond *t + 1*
  - Davis et al. (2012) and Li (2010) → 3 quarters ahead for quarterly tone
  - Alves et al. (2016) → 2 years ahead for annual tone

SESSION 3: ANNUAL REPORTS & EARNINGS ANNOUNCEMENTS

Information *cont.*

- Do narratives have incremental information content?

$$CAR_{it} = \lambda_0 + \lambda_1 UE_{it} + \lambda_2 \Delta Tone_{it} + \sum_{k=1}^{K} \phi_k Controls_{kit} + v_{it}$$

- Consistent evidence using US data that $\hat{\lambda}_2 > 0$ for:
  - Earnings announcement press release → 3-day window centred on earnings announcement (Davies et al. 2012, Henry & Leone 2016)
  - MD&A section of annual report on Form 10-Q and 10-K → 3-day window centred on the SEC filing date

- Also evidence that narratives mitigate market mispricing
  - Li (2010) finds that management discussion about implications of accruals for future performance reduces accruals mispricing

- Conclusion → narratives are informative and investors understand this

SESSION 3: ANNUAL REPORTS & EARNINGS ANNOUNCEMENTS

Obfuscation

- Do preparers use narratives to disguise unfavourable performance?

- Yes! → evidence supports obfuscation (impression management)
  - 10-K text is more complicated/harder to read (high Fog index) when current performance is poor and performance is less persistent (Li 2008)
  - Increasing demand for analyst services for firms with less readable 10-Ks (Lehavy et al. 2011)
  - Unusually high optimism in earnings announcements predicts lower future performance and impression management behavior (Huang et al. 2014)
  - MD&A commentary characterized by managerial self-attribution bias → (un)favourable outcomes attributed to (external) internal factors (Li 2010)

- Critique
  - Surprising if obfuscation wasn't evident → where's the tension?
  - Similar conclusions with manual methods (Merkl-Davies & Brennan 2011)

SESSION 3: ANNUAL REPORTS & EARNINGS ANNOUNCEMENTS

Note of caution

- Researchers tested information and obfuscation hypotheses using smaller samples based on manual coding
  - Insights and conclusions not dissimilar to more recent large-sample studies → evidence to support both perspectives (Garcia Osma & Guillmon-Saorin 2011)

- Understand the comparative advantage(s) of large-sample textual analysis to ensure contribution threshold is achieved
  - Greater objectivity and replicability
  - Greater generalizability
  - Higher statistical power → depends

  May not be sufficient unless very carefully motivated

  - Measure aspects of text or test predictions that would be difficult otherwise

  → Research question still fundamental determinant of contribution → then match methodology to question

  → Rennekamp (2013) uses experiment to examine impact of readability on investor reaction

SESSION 3: ANNUAL REPORTS & EARNINGS ANNOUNCEMENTS

Richer insights on informativeness

- Rather than focusing on average effects, a body of work seeks to identify conditioning variables → factors affecting the degree of informativeness
  - Better fit with automated textual analysis methods because (relatively) large samples required to test for intervening effects

- Insights include:
  - MD&A modifications drive reaction to 10-K filing (Brown and Tucker 2011)
  - Credibility influences the extend to which soft information is influential in the price formation process (Demers & Vega 2015)
  - *ΔTone* predicts returns in 2-day window after SEC filing date where information environment is weak (Feldman et al. 2013)
  - Net optimism more useful where earnings less informative about firm value
  - MD&A content explains firm value where accounting less useful (Ball et al. 2014)
  - Forward-looking MD&A disclosures more useful where stock prices have low informational efficiency, particularly for loss firms (Muslu et al. 2015)

SESSION 3: ANNUAL REPORTS & EARNINGS ANNOUNCEMENTS

Beyond the US

- Large-sample evidence for textual analysis is limited outside the US
- International evidence (Lang & Stice-Lawrence 2015)
  - Text attributes such as length, boilerplate and complexity are predictably associated with regulation and incentives for more transparent disclosure
  - Improvements in annual report disclosures associated with improvements in economic outcomes → liquidity, institutional ownership, analyst following
- UK evidence (Alves et al. 2016)
  - Significant variation across firms and time in the way firms present narrative information → disclosure is about structure/presentation as well as content
  - Disclosure changes associated with IFRS adoption limited to the financial statements component of the annual report
  - Performance-related disclosures predict future earnings but effects vary with preparers' incentives

SESSION 3: ANNUAL REPORTS & EARNINGS ANNOUNCEMENTS

Consistency matters

- Davis & Tama-Sweet (2012) examine differences in optimistic language between the earnings announcement and the MD&A
  - More optimistic language in earnings press releases because earnings announcements are associated with larger price responses
  - Level of incremental optimism reflects preparers' reporting incentives

- Dikolli at al. (2014) compare language in letter to shareholders (unregulated) with language in the MD&A (regulated)
  - Differences shed light on CEO credibility and integrity

- Alves et al. (2016) compare net tone in chairman's letter (outsider-authored) with the MD&A (insider-authored)
  - Unusually positive MD&A language (relative to chairman's letter) associated with lower predictability → inconsistency reflects obfuscation

SESSION 3: ANNUAL REPORTS & EARNINGS ANNOUNCEMENTS

Capturing latent constructs

- Use annual report and earnings announcement disclosures to measure firm- and manager-specific characteristics and study their impact
  - Good fit with automated textual analysis methods because sample size and replicability are critical

- Measures derived from form 10-K include:
  - Risk (Campbell et al. 2013, Kravet & Muslu 2013)
  - Competition (Li et al. 2013, Bushman et al. 2015)
  - Business strategy (Kabanoff and Brown 2008)
  - Fraud risk (Purda & Skillicorn 2015)
  - Financing constraints (Bobnaruk et al. 2015)
  - CEO integrity (Dikolli et al. 2014)
  - Trust in corporate culture (Audi et al. 2014)

SESSION 3: ANNUAL REPORTS & EARNINGS ANNOUNCEMENTS

The annual report

- Surprisingly little known about the annual report despite central and enduring role in the corporate communication process
  - Large sample automated textual analysis offers scope for new insights

- Dyer et al. (2016) examine properties of 10-K and changes therein over time → Why is the 10-K getting longer and does it matter?
  - FASB and SEC compliance requirements account for much of the expansion and drive increases in complexity and redundancy
  - Only risk factor disclosures appear useful to investors

- Annual report on Form 10-K not representative of annual reports filed by non-US registrants
  - More discretion given to preparers outside the US in terms of structure, presentation and content
  - Opportunity to examine new dimensions of disclosure

TRIPLE-ACCREDITED, WORLD-RANKED

Lancaster University Management School

Narrative information in financial markets

Session 4

Evidence: Other sources of information

Steven Young

(Lancaster University)

6th WHU Doctoral Summer Program in Accounting Research

Current Issues in Empirical Financial Reporting Research

11-14 July, 2016

Session 4: Other Documents

Overview

- Other regulatory filings and corporate announcements
  - Prospectuses, 8-Ks, press releases, etc.
- Conference call transcripts
- Analyst reports
- Financial media articles
- Social media and blogs
  - Twitter, SeekingAlpha, etc.
- Regulatory pronouncements and speeches
- Other options (comment letters, corporate websites, crowdfunding sites)

SESSION 4: OTHER DOCUMENTS

Session objectives

- Review research on textual disclosures associated with other elements of corporate information environment
- Consider the research benefits associated with different sources of textual commentary
- Highlight potential gaps in extant research

SESSION 4: OTHER DOCUMENTS

Other regulatory filings and corporate announcements

- Firms required to comply with a range of filing regulations, majority of which involve textual content

- Example 1 → Form S-1 (and Form 424) language and IPO pricing (Loughran & McDonald 2014 *JFE*)
  - Form S-1 is the first SEC filing in the initial public offering (IPO) process
  - Definitiveness of language on business strategy and operations used as a proxy for ex ante uncertainty regarding valuation
  - Less definitive language (more uncertainty) → higher first-day returns
  - Evidence supports theoretical models predicting positive link between ex ante valuation uncertainty and initial returns (Beatty & Ritter 1986)
  - Language-derived measure of uncertainty explains underperformance better than many traditional IPO controls

SESSION 4: OTHER DOCUMENTS

Other regulatory filings and corporate announcements *cont.*

- Example 2 → Form 8-K announcements and stock price prediction (Lee et al. 2014)
  - Use financial events reported in 8-Ks to forecast the direction of future stock price changes → up, down, no change
  - Words lemmatized and converted to unigram features using Pointwise Mutual Information (PMI)
  - Combine unigram features in a random forest classifier
  - Linguistic features have incremental predictive ability for one-day-ahead returns, with weaker predictive ability for returns up to 5 days ahead
  - Publish corpus → http://nlp.stanford.edu/pubs/stock-event.html

SESSION 4: OTHER DOCUMENTS

Other regulatory filings and corporate announcements *cont.*

- Example 3 → Corporate press releases during merger negotiations (Ahern & Sosyura 2014)
  - Bidders in stock mergers issue an abnormally high number of news stories after start of merger negotiations but prior to public announcement
  - Fixed ratio bidders issue fewer *negative* stories (based on L&M dictionary) during period when exchange ratio is established
  - Fixed ratio bidders also use fewer *understated* words (Harvard IV dictionary) during negotiation period → but no more likely to *overstate*
  - Price correction observed for fixed ratio bidders after the merger announcement
  - Overall, evidence consistent with management engaging in active media strategy → "spin"

SESSION 4: OTHER DOCUMENTS

Other regulatory filings and corporate announcements *cont.*

- Example 4 → Management earnings forecasts (Baginski et al. 2011)
  - Sentiment (based on L&M dictionaries) directionally consistent with the quantitative earnings forecast
  - Sentiment more informative when past earnings less informative for valuation
  - Sentiment pricing lower for richer pre-disclosure information environment
  - Negative sentiment associated with higher stock return volatility, although effect is attenuated when the press release contains more uncertain language

SESSION 4: OTHER DOCUMENTS

Other regulatory filings and corporate announcements *cont.*

- Example 4 → Management earnings forecasts (Baginski et al. 2011)
  - Sentiment (based on L&M dictionaries) directionally consistent with the quantitative earnings forecast
  - Sentiment more informative when past earnings less informative for valuation
  - Sentiment pricing lower for richer pre-disclosure information environment
  - Negative sentiment associated with higher stock return volatility, although effect is attenuated when the press release contains more uncertain language

- Example 5 → Language in chairman's letters (Craig et al. 2013)
  - Letters from Chair of Indian firm Satyam (2002-2008) → fraud revealed 2009
  - Personal pronouns, tone, extreme emotion, and Diction's *CERTAINTY* construct
  - Strong use of first-person plural pronouns, dominant positive tone, and extreme positive emotion in run-up to fraud revelation
  - Deceivers divert suspicion by reducing direct references to themselves and by increasing the positive tone and degree of extreme positive emotion

SESSION 4: OTHER DOCUMENTS

Conference call transcripts

- Conference call transcripts provide a rich venue for large-sample analysis
  - Overview presentation delivered by senior management → scripted
  - Q&A session between management and analysts → more spontaneous

- Example 1 → Manager-specific tone in conference calls (Davis et al. 2015)
  - Net tone measured using wordlists from Diction, Henry and L&M
  - Focus on component of conference call tone not explained by current and future performance or strategic incentives → presentation plus Q&A sections
  - Manager-specific fixed effect explains 6-7% of variation in residual tone
  - Manager-specific tone correlates with observable factors including gender, career incentives and involvement in charities → cognitive characteristics

  ⇒ Manager tone reflects individual's tendency for optimism
  - Weak effect of manager fixed effect on announcement-period returns → choice of language impacts market response

Session 4: Other Documents

Conference call transcripts *cont.*

- Example 2 → Incremental content of Q&A (McKay Price et al. 2012)
  - Call tone (Henry and Harvard IV-4) for Q&A, controlling for presentation
  - Tone explains (predicts) announcement-period returns and trading volume
  - Tone also explains 60 trading-day post-earnings announcement drift → more important than earnings surprise
  - Results more pronounced for context-specific (Henry) dictionary

Session 4: Other Documents

Conference call transcripts *cont.*

- Example 2 → Incremental content of Q&A (McKay Price et al. 2012)
  - Call tone (Henry and Harvard IV-4) for Q&A, controlling for presentation
  - Tone explains (predicts) announcement-period returns and trading volume
  - Tone also explains 60 trading-day post-earnings announcement drift → more important than earnings surprise
  - Results more pronounced for context-specific (Henry) dictionary

- Example 3 → Predicting restatements (Larcker & Zakolyukina 2012
  - Linguistic-based predictors of deceptive discussions
  - Deceptive management use more general knowledge references, more nonextreme positive emotion words, fewer references to shareholder value
  - Deceptive CEOs use more extreme positive emotion and fewer anxiety words
  - Portfolio of firms with the highest deception scores from CFO narratives generate annualized alphas between -4% to -11%

SESSION 4: OTHER DOCUMENTS

Analyst reports

- Analyst reports provide offer a unique window on financial statement analysis and the valuation process

- Example 1 → Information content (Huang et al. 2014)
  - Do discussions in analyst reports contain incremental information beyond earnings and price forecasts, and investment recommendation?
  - Extract sentence-level opinions (positive, negative, neutral) using naïve Bayes classifier → aggregate sentence-level opinions to produce report-level opinion
  - One stdev. increase in favourableness of textual opinion results in an additional 2-day abnormal return of 41 bp
  - Market reacts to quantitative summary measures more intensely when accompanying textual opinion is confirmatory
  - Reaction is 2× larger for negative opinion than positive opinion
  - Incremental predictive value for future earnings growth up to 5 years out

SESSION 4: OTHER DOCUMENTS

Analyst reports *cont.*

- Example 2 → Report readability (De Franco et al. 2015)
  - Readability measured using Fog Index
  - High ability analysts issue more readable reports
  - Trading volume higher for more readable reports → consistent with theory predicting more precise information leads to more trading

SESSION 4: OTHER DOCUMENTS

Financial media articles

- The business press represents a potentially rich (objective?) perspective on financial performance
  - Extant research already demonstrates how the financial media can add value to the reporting process (Miller 2006, Drake et al. 2014)
  - Media articles obvious target for textual analysis applications

- Example 1 → News sentiment predicts performance (Tetlock et al. 2008)
  - *WSJ* and *DJNS* stories for S&P 500 firms from 1980-2004
  - Fraction of negative words (Harvard IV-4) in firm-specific news stories
  - Negative sentiment predicts earnings beyond past earnings and forecasts
  - Predictive ability for short-run (1-day) returns → markets briefly underreact to negativity
  - Predictive ability for earnings and returns is more pronounced for articles focusing on fundamentals → reference to word stem "earn"

SESSION 4: OTHER DOCUMENTS

Financial media articles *cont.*

- Example 2 → Does the dictionary matter? (Heston and Sinha 2014)
  - Sentiment measured using Harvard IV-4 and L&M dictionaries, plus Thomson-Reuters neural network classifier (positive, negative, neutral)
  - Sophisticated sentiment measure predicts larger and more persistent returns
  - Aggregating sentiment over 1 week predicts returns up to 13 weeks ahead
  - Prices react more quickly to positive news relative to negative news

SESSION 4: OTHER DOCUMENTS

Financial media articles *cont.*

- Example 2 → Does the dictionary matter? (Heston and Sinha 2014)
  - Sentiment measured using Harvard IV-4 and L&M dictionaries, plus Thomson-Reuters neural network classifier (positive, negative, neutral)
  - Sophisticated sentiment measure predicts larger and more persistent returns
  - Aggregating sentiment over 1 week predicts returns up to 13 weeks ahead
  - Prices react more quickly to positive news relative to negative news

- Example 3 → More advanced NLP applications (Malo et al. 2013)
  - Focus on phrase structure and semantic orientation rather than individual words and sentences → overall meaning can differ from word-level polarity
  - Traditional sentiment measures (e.g., movie reviews) based adverbs and adjectives → financial sentiment involves language of *expectations*
  - Present human-annotated financial phrase-bank
  - Develop a new Linearized Phrase Structure (LPS) model of semantic orientation to measure sentiment in financial text

SESSION 4: OTHER DOCUMENTS

Social media and blogs

- Social media, blogs, message postings, etc. offer insight on sentiment
  - Facebook exploit social media posts to construct *Gross Happiness Index*

- Example 1 → Measuring sentiment and investor type (Das & Chen 2007)
  - Extract retail investor sentiment from postings on stock message boards
  - Use various classifiers to measure sentiment
  - Useful for assessing impact on investor opinion of earnings announcements, press releases, regulatory changes, third party news

- Example 2 → Predictive power of social media opinions (Chen et al. 2014)
  - Articles and comments on social media platform for investors → Seeking Alpha
  - Sentiment measure → fraction of negative words (L&M dictionary)
  - Negativity predicts returns (up to 3 months out) and earnings surprises
  - Reason(s) for predictive power is unclear

SESSION 4: OTHER DOCUMENTS

Social media and blogs *cont.*

- Example 3 → Private information using social media (Bok et al. 2016)
  - Messages from Twitter → distinguish between local and nonlocal Twitter users
  - Negative tone (L&M dictionary) of local tweets predicts future stock returns and earnings → no predictive ability for nonlocal tweets
  - Local social media activity reflects new information
  - Negative tone of local tweets leads to higher bid-ask spreads and lower depths
  - Sharing information within individuals' network *increases* information asymmetry among investors

SESSION 4: OTHER DOCUMENTS

Regulatory pronouncements

- Announcements by financial market regulators and economic policy are commonplace and economically significant

- Example 1 → Characteristics and impact of European Council communications (Wisniewski & Moro 2014)
  - Extend literature on links between politics and finance by applying NLP techniques (General Inquirer dictionaries)
  - Positive sentiment (positive words) → positive price impact
  - Obfuscation (abstract vocab.) → negative market impact
  - Statements of moral integrity (rectitude gain words) → positive price reaction

- Example 2 → Extracting financial fraud information from SEC litigation releases (Zaki & Theodoulidis 2013)
  - Use text mining methods to extract characteristics litigation release number, release publication dates, manipulation participants, timeline, actions

SESSION 4: OTHER DOCUMENTS

Regulatory pronouncements *cont.*

- Opportunities → Hendricks et al. (2015) examine how firms respond to *proposed* regulation on Mortgage Service Rights associated with Basel III
  - MSR only one aspect of proposals; relevance likely to vary across firms
  - Comment letters could reveal importance of MSR proposal for individual banks

SESSION 4: OTHER DOCUMENTS

Other text resources

- Corporate (IR) websites contain a wealth of narrative information that to date has remained largely unexplored on a large sample basis
  - Webpages are dynamic → no systematic archiving
- Business websites where narratives are central to business model
  - Embryonic work in the area of crowdfunding applying NLP techniques to understand funding outcomes (Gao & Lin 2015)
- Comment letters / lobbying activity
  - Unstructured format plus redundant content (e.g., address) creates challenges
- Financial market regulations, banking rules, accounting standards, etc.
  - Use corpus/NLP methods to study evolution of rules, consistency of themes...
- Emails → Louwerse et al. (2010) study Enron email dataset