Online Appendix
for
What are You Saying? Using topic to Detect Financial Misreporting

Nerissa C. Brown
Associate Professor
University of Illinois
nerissab@illinois.edu

Richard M. Crowley
Assistant Professor
Singapore Management University
rcrowley@smu.edu.sg

W. Brooke Elliott
Ernst & Young Distinguished Professor
University of Illinois
wbe@illinois.edu

November 2019
Appendix A.1 10-K Parsing

To parse the 10-K filings, we implement a layered approach using the Natural Language Toolkit (NLTK), HTMLParser, and regular expressions in Python (2.7). We derive several of our steps from Li [2008] and mark these with an asterisk (*).

We start by filtering out any auto-generated electronic filings. Such filings occur when a firm submits a paper filing and can be easily identified by the following statement: “This document was generated as part of a paper submission.” Next, we remove large portions of the documents that are not written paragraphs by leveraging regular expressions in Python. We strip out all XBRL data, which account for a significant portion of the file size in later years, but do not contain any plain text that is not otherwise included in the body of the document. We detect XBRL data by searching for XBRL INSTANCE in any letter case (i.e., lower or upper case, etc.) as well as ex-101 INSTANCE DOCUMENT, also in any letter case. We further remove any outlier text appearing after the </TEXT> tag, which indicates the end of the 10-K.

The next step removes all forms of text formatting which can create processing issues. We remove all paragraph tags (<p>) while retaining the paragraph structure of the document. We replace nonbreaking spaces (&nbsp and &#160) with standard spaces, as some firms use these in place of standard spaces. We also filter out characters by removing bullets (e.g., &#149, &#8226, •, and $bull;) and other punctuation used to create bullets. We identify these other punctuation by removing certain punctuation marks that are preceded by whitespace, including the following characters in parentheses: (,.?!*+).

From the remaining text, we remove the SEC header by trimming any text before the <TEXT> tag.*1 We then remove all tables by deleting any text between the tags <TABLE> and </TABLE> (in any letter case).* Lastly, we delete all other HTML tags* (which we remove with NLTK), all images embedded by EDGARwiz, and all other hexadecimal data embedded

---

*1While this method differs slightly from Li [2008], it has the same effect as trimming between <SEC-HEADER> and </SEC-HEADER>, but removes a greater amount of excess metadata around the header tags.
in the files. The EDGARwiz images typically begin with GRAPHIC and end with an accent grave (‘), followed by spacing, newlines, and the word end.

We begin to process the text into a machine-readable form using regular expressions in Python. Our program adjusts multiple spaces into single spaces, removes leading and ending spaces, changes multiple newlines to single newlines, deletes page numbers (along with characters surrounding them, including the following characters: ‘-_.=+:~’), and removes fixed width enters (to rebuild the paragraph structure of a fixed width file). The program deletes all paragraphs that are less than 50% alphanumeric and fixes odd hyphenation (e.g., best-in-class to best-in-class). After these steps, we use HTMLParser to convert any HTML encoded characters to machine-readable form (e.g., &amp; to &amp;) and then encode the files as ASCII.

At this point, we capture the MD&A statement (if present), by searching for a large number of phrases with the following structure: [beginning of a line] [optional whitespace] [first phrase] [1,000 to 250,000 characters] [beginning of a line] [optional whitespace] [second phrase]. We include a character restriction to avoid conflicts with phrases across the document, such as an introduction. For the first phrase, we use the following (in any letter case, with optional characters in square braces):

- management[‘]s discussion
- item [whitespace] 6 [fewest number of characters possible] management[‘]s discussion
- item [whitespace] 7 [fewest number of characters possible] management[‘]s discussion
- item [whitespace] 8 [fewest number of characters possible] management[‘]s discussion
- item [whitespace or periods] management[‘]s discussion

For the second phrase, we use the following (in any letter case, with optional characters in square braces):

- financial statements
Lastly, we remove all paragraphs with < 80 characters. This step primarily captures short headings, tables of contents, and other short parts of the document that are unlikely to be proper paragraphs.
Appendix A.2 Simulation

We determine the stopword parameters and the number of topics for the LDA algorithm to discover by running a short simulation to optimize the model for in-sample performance. We seeded the simulation with initial stopword parameters of (50%, 500, and 100) and 20 topics to identify. The simulation was run parameter by parameter (starting with the first stopword parameter) rather than as a four-dimensional optimization problem, as the latter method (while more accurate) is computationally infeasible. For each parameter, we run the simulation twice, first to pinpoint possible locations of the optimal parameter and second to more precisely pinpoint the optimal parameter value.

For each pass of the simulation, we randomly select 1,000 filings per year, parse the selected filings using the stopword parameters, run the LDA algorithm, and then estimate each logistic regression to determine the in-sample statistical power of the respective specification (measured as the pseudo \( R^2 \) of the logistic regression). We conduct this analysis for each five-year window in the main sample. The parameter that resulted in the highest mean pseudo \( R^2 \) was selected as the optimal parameter. With respect to the second and third stopword parameters, we made a slight adjustment to the mean to control for the lack of convergence, as higher values of these parameters are likely to lead to convergence problems for our small sample of documents. The adjustment uses a smoothed average of the pseudo \( R^2 \) values for nearby parameters in a given year when a specification did not converge. The results from our simulations indicate optimal stopword parameters of (60%, 1100, 100) and 31 topics as the optimal identification number for the LDA algorithm. Figure A.1 depicts graphs of the average pseudo \( R^2 \) values.
This chart shows the results of the simulation exercise described in Appendix A.2.
Appendix A.3 List of Combined Topics

The table below provides a list of the combined topics identified across our sample period. The combined topics are derived by grouping the individual LDA topics across years based on the Pearson correlation of the word weights within the topics. All topics with a Pearson correlation above a set threshold are grouped together, resulting in 64 combined topics. For each combined topic, we present the topic number and label (both in bold) along with 10 representative bigrams and 2 representative sentences. The topic labels are based on interpretations of the top 20 bigrams from a set of 334 representative sentences for each topic. The bigrams presented are the 10 most common n-grams after excluding redundant phrases (e.g., “compared with” and “compared to” in topic 2). The sentences presented are from the 334 representative sentences per topic and are drawn from the set of top 100 sentences with the highest cosine similarity to the other representative sentences. We place all industry-specific topics at the end of the list (topics 50 through 64) for ease of exposition.

| 1) Decrease in income compared to prior periods: | compared to, gross profit, other income, company contributed, operating income, company expects, gross margin, income decreased, capital expenditures, decreased to |
| Management fee income decreased in 1998 to $0 as compared to $1.4 million in 1997. |
| The Company’s gross profit margin decreased to 59% in Fiscal 1996, compared to 65% in Fiscal 1995. |

| 2) Increase in income compared to prior periods: | compared with, gross margin, income was, operating income, gross profit, other income, fiscal compared, income taxes, non-interest income, profit was |
| Operating profit was $122.8 million in 2011, compared with $113.9 million in 2010, an increase of 7.8%. |
| Other revenue in 2010 was $27,000, an increase of 69% compared to $16,000 in 2009. |

| 3) Postretirement health care benefits assumptions: | health care, care plans, assumed health, trend rates, care cost, cost trend, effect on the amounts, rates have, significant effect, postretirement health |
| Assumed health care cost trend rates have a significant effect on the amounts reported for the health care plans. |
| The health care cost trend rate is the assumed rate of increase in per capita health care charges |

| 4) Fuel costs and commitments: | nuclear fuel, fuel related, coal mining, coal reserves, fuel costs, fuel expense, fuel supply, fuel commitments, related expenses, fossil fuel |
| Fuel costs are a significant portion of the Company’s total operating expenses. |
| Fuel commitments consist of commitments for nuclear fuel coal and coal transportation costs. |

| 5) Nuclear waste disposal costs: | nuclear decommissioning, nuclear power, nuclear fuel, spent nuclear, nuclear plant, decommissioning costs, nuclear waste, waste policy, disposal of spent, nuclear business |
| APS recorded $11 million for nuclear decommissioning expense in each of the years 2002, 2001 and 2000. |
| The NRC requires nuclear power plant owners to annually update minimum financial assurance amounts for the future decommissioning of nuclear facilities. |

Continued on next page
6) **Financial statement information**: dollars in millions, ended december, income taxes, financial statements, accompanying notes, millions except, year ended, pension benefits, consolidated balance, consolidated financial

The components of net deferred tax assets and liabilities are as follows in millions.

All amounts are in millions of dollars except for the earnings per share data.

7) **Restaurant business growth**: company operated, operated restaurants, company owned, franchised restaurants, company opened, operated restaurants, restaurants at december, franchisees opened, owned and operated, opened restaurants

At December 31, 2001, the Company’s system included 821 restaurants, comprised of 235 company-owned and operated restaurants and 586 franchised restaurants.

The Company opened 28 new restaurants, acquired two restaurants from franchisees and closed four restaurants in 2001.

8) **Derivatives and hedging**: derivative financial, financial instruments, trading purposes, derivative instruments, instruments for trading, hold or issue, issue derivative, enter into, hedging activities, speculative purposes

The Company does not hold or issue derivative financial instruments for trading purposes.

The Company has not used derivative instruments for commodity hedging purposes.

9) **Real estate loan operations**: real estate, national bank, estate loans, estate investment, loan bank, home loan, federal home, investment trust, trust REIT, mortgage loans

The Bank emphasizes multifamily and commercial real estate lending.

The Bank provides a variety of real estate loans secured by residential real estate.

10) **Accounts receivable and doubtful accounts**: accounts receivable, doubtful accounts, allowances for doubtful, valuation allowances, vendor allowances, product returns, allowances include, estimated allowances, uncollectible amounts, based on historical

Tenant and other receivables are recorded net of the allowances for doubtful accounts.

The allowances for future product returns are reflected as a reduction of revenue and accounts receivable.

11) **Corporate spin-offs**: prior to the spin, spin off from, spin off transaction, financial statements, since the spin, adjusted to reflect, completed the spin, after the spin, common stock, been adjusted

The Company is currently in the process of completing all of the applicable steps necessary to effect the spin off transaction.

In anticipation of the spin off the Company was incorporated on March as a Delaware corporation.

12) **Segment performance**: generation segment, discussed below, segment as discussed, because of items, segment because, merchant services, other operations, increased million, items detailed, maintenance expenses

Other operations and maintenance expenses increased million in the Merchant Generation segment as discussed below.

This increase was primarily due to the impact of growth in our International merchant services segment.

13) **Foreign country risks**: republic of china, united states, located in china, agreement with, entered into, future inflation, inflation in china, conduct business, business in china, china may inhibit

Our investments in China create risks and uncertainties relating to the laws in China.

Future inflation in China may inhibit our ability to conduct business in China.

14) **Laser products**: excimer laser, laser system, laser vision, laser technology, laser printers, laser based, laser beam, laser products, capital expenditures, discontinued operations

The excimer laser is the dominant laser used for the treatment of refractive disorders.

The Company sells primarily to customers involved in the application of laser technology and the manufacture of telecommunications products.

15) **Digital technology and services**: digital media, internet access, high speed, digital signal, analog and digital, internet services, services include, cable television, digital imaging, internet based

We are in the process of developing mixed signal and digital technology to decode digital signals.

The Company holds numerous digital patents and relationships with key manufacturers and suppliers in digital technology.

16) **Long term assets**: property and equipment, equipment property, property plant, stated at cost, intellectual property, equipment consisted, carried at cost, long lived, intangible assets, assets include

Property and Equipment Property and equipment are stated at cost.

Property and equipment are stated at cost.

Continued on next page
17) **Franchise revenue recognition**: franchise fees, franchise agreements, initial franchise, franchise royalties, franchise operations, franchise revenues, franchise rights, development fees, intangible assets, brand name

Upon the expiration of the term of a franchise, the Company may apply for a franchise renewal.

Franchise fee revenue is recognized upon opening of the franchise store.

18) **Business structure**: holding company, bank holding, loan holding, company under, financial holding, holding corporation, utility holding, holding companies, unrealized holding, holding gains

The Company is registered as a bank holding company under the Bank Holding Company Act of 1956.

On July 11, 2001, the Company completed its reorganization into a holding company structure.

19) **Debt issuance**: convertible debenture, common stock, subordinated debenture, agreement with, purchase agreement, principal amount, note or debenture, debenture holders, company issued, debenture offering

As part of the strategic alliance, BNFL invested $10.0 million in the Company in the form of a convertible debenture.

The increase is primarily due to the sale of the convertible debenture and the bridge financing.

20) **Media and entertainment**: interactive entertainment, entertainment software, entertainment company, entertainment group, agreement dated, entertainment industry, entertainment services, company acquired, lease agreement, digital entertainment

Digital entertainment systems offer the DVD+RW drives as well as digital entertainment center products.

We also introduced new entertainment products including online entertainment services and magazine subscriptions.

21) **Food products and services**: food service, drug administration, food and drug, food products, food packaging, food and beverage, food processors, food distribution, quality food, food processing

Our purchasing department administers programs for the procurement of food and non-food products.

Food accounted for 61% of total food and general merchandise sales in 2002 and 2001.

22) **Merger activities**: merger with, merger agreement, plan of merger, prior to the merger, completed a merger, merger related, agreement and plan, proposed mergers, approved the merger, closing of the merger

The reverse merger was treated as a re-capitalization of the Company.

The increase in overall EBITDA primarily relates to the impact of the Merger.

23) **R&D partnerships**: pharmaceutical products, collaboration agreement, agreement between, pharmaceutical company, collaboration with, entered into, pharmaceutical services, license agreement, research collaboration, between the company

In 2000, Molecumetics entered into a collaboration agreement with Athersys for the development of small molecule drug candidates.

The GlaxoSmithKline collaboration agreement and many of our other collaboration agreements will expire in 2001.

24) **Consolidated financial information**: consolidated statements, cash flows, statements of cash, subsidiaries consolidated, corporation consolidated, consolidated balance, balance sheets, comprehensive income, share amounts, company consolidated

Amounts recognized in the consolidated statements of stockholders equity and comprehensive income in millions.

Consolidated Statements Of Income in millions, except per share amounts.

25) **Advertising expenses**: advertising costs, company expenses, advertising expense, costs are expensed, expensed as incurred, first time, costs as incurred, online advertising, advertising revenue, advertising and promotion

The Company expenses advertising and promotional costs as incurred.

Advertising costs are expensed as incurred and were not material for 2007, 2006 and 2005.

26) **Country risks**: country to country, from country, united states, vary from, country basis, country by country, based on the country, outside the united, country risk, widely from

The relative importance of government and private systems varies from country to country.

The Firm has a comprehensive process for measuring and managing its country exposures and risk.

27) **Joint venture agreements**: joint venture, joint and several, full and unconditional, entered into, venture agreement, agreement with, venture between, venture partners, joint plan, formed a joint

The joint venture customer partner, ADE, owns 15% of the joint venture.

The Company assumed the debt of the joint venture of $9.7 [Million].

Continued on next page
28) **Credit card arrangements**: credit card, debit card, card transactions, card receivables, card services, card processing, gift card, card loans, interchange fees, card revenue, 

Compensation costs related to the credit card agreements are recorded as contra-revenue in card income.

Credit card fees increased $814,000 primarily due to a higher volume of credit and debit card transactions.

29) **Fair value/cash flow hedging**: interest rate, rate swap, swap agreement, entered into, notional amount, fair value, floating rate, swap transactions, currency swap, hedge accounting

The rate on the interest rate swap agreement covering $155.3 million of the 2.25%.

At December 31, 2005, the notional value of the interest rate swap was $5.5 million.

30) **Purchase and acquisition agreements**: common units, each purchaser, authorized purchaser, units purchased, agreement with, with the purchaser, entered into, third party, affiliated purchaser, place an order

As part of the agreement to sell the net assets of STMontreal, the Company released the purchaser from this guarantee.

The Company provides a one year warranty from the date of shipment to the original purchaser.

31) **Legal proceedings**: district court, southern district, district of new york, bankruptcy court, court of appeals, northern district, court has not yet ruled, eastern district, supreme court, appeals for the district

These class action suits were removed from state court and transferred to the U.S. District Court for the Southern District of California.

We and a number of other major pharmaceutical and biotechnology companies are named defendants in certain Average Wholesale Price litigation pending in the U.S. District Court for the District of Massachusetts alleging, among other things, violations in connection with Medicaid reimbursement.

32) **Strategic alliances**: strategic alliance, alliance with, entered into, agreement with, alliance agreement, alliance partners, into a strategic, company entered, license agreement, ventures sold

As part of the spin-off, Sunburst and the Company entered into a strategic alliance agreement.

The Company recognized the research funding portion of the alliance on a percentage of completion basis.

33) **Credit agreements**: jpmorgan chase, kinder morgan, vice president, agreement with, entered into, credit facility, facility with, chase manhattan, credit agreement, company entered

On November 1, 2005, the Company amended its $25,000,000 credit facility to change the financial metrics that must be met to remain in compliance with the debt covenants for the third and fourth quarters of 2005 and thereafter, to accommodate the Company’s revised financial outlook.

On November 21, 2005, SS/L entered into the $20,000,000 letter of credit agreement with JPMorgan Chase Bank.

34) **Patent infringement and rights**: patent applications, patent and trademark, trademark office, regulations provide, patent infringement, procedure for challenging, patent rights, control presumption, rebuttable control, filed a patent

During fiscal 2005, the Company acquired a license for the rights to a patent for $2.0 million.

The Company has filed an answer and counterclaim alleging invalidity of the patent.

35) **Stock option plans**: board of directors, stock option, shall not exceed, this agreement, shall be entitled, company shall, shall be entitled, option shall, meaning set forth, shall become

The term of the options shall not exceed ten years from the date of grant.

Stock options granted pursuant to the Plan shall be authorized by the board of directors.

36) **Share capital**: preferred stock, redeemable preferred, mandatorily redeemable, cumulative redeemable, investing activities, convertible preferred, series A cumulative, common stock, series A preferred, capital securities

The 2004 amounts are recorded as a part of the class A redeemable preferred stock balance.

Redeemable preferred stock and Common Stock warrants decreased as a result of the 2005 Recapitalization.

37) **Partnership arrangements**: general partner, limited partner, sole general, managing general, operating partnership, managing partner, limited partnerships, partner interest, executive officers, responsible for managing

Our general partner owns the 2% general partner interest and all of the incentive distribution rights.

Our capital accounts are comprised of a 2% general partner interest and 98% limited partner interests.

Continued on next page
38) **Equity ownership and control**: common stock, consolidated statements, property trust, insurance company, agreement with, limited partnership, property limited, life insurance, year ended, december ended

On August 25, 2005, the Company acquired all of the outstanding capital stock of Lincoln Technologies Inc., which was accounted for as a purchase under SFAS No. 141.

On January 16, 2007, Liberty completed the sale of its controlling interest in OpenTV Corp (“OPTV”) to an unaffiliated third party for cash consideration of $132 million, $20 million of which was deposited in an escrow account to fund potential indemnification claims by the third party made prior to the first anniversary of the closing.

39) **Share capital transactions**: real estate, preferred stock, convertible preferred, estate loans, commercial real, series A convertible, series A preferred, preferred units, estate construction, construction loans

The Company received proceeds of $17.68 million for the Series A Preferred Stock and the Warrant.

The investment consists of 19,000 shares of Series A Convertible Preferred Stock.

40) **Segment performance**: segment information, operating segment, business segment, segment consists, operating income, segment performance, performance based, reportable segment, segment reporting, segment results

The increase was primarily due to the gross profit percentage increase of 0.3% by the U.S. Distribution segment and the decrease of 0.5% in the Europe segment.

The segment contributed approximately 32% of the Company’s total segment profit.

41) **Securitized/guaranteed securities**: guaranteed securities, fannie mae and freddie, mortgage backed, backed securities, farmer mac guaranteed, mortgage loans, preferred stock, government sponsored, securities issued, freddie mac preferred

Fannie Mae and Freddie Mac mortgage related securities are guaranteed as to payment of principal and interest by the respective entity issuing the security.

Recent capital restructuring at Fannie Mae and Freddie Mac and developments in the residential mortgage business have resulted in impairment of these securities.

42) **Environmental risks**: mining operations, environmental liabilities, environmental remediation, environmental regulation, environmental risks, environmental laws, environmental compliance, environmental protection, environmental costs, mining claims

Environmental compliance and remediation costs and the costs of environmental liabilities could exceed our estimates.

Recent capital restructuring at Fannie Mae and Freddie Mac and developments in the residential mortgage business have resulted in impairment of these securities.

43) **Foreign currency risks**: functional currencies, foreign currencies, local currencies, asia pacific, foreign subsidiaries, denominated in foreign, company’s foreign, foreign operations, consolidated statements, respective local

The functional currencies of the Company’s foreign subsidiaries are primarily the respective local currencies.

A portion of our costs and expenses are denominated in foreign currencies.

44) **Geographic locations**: located in, united stated, primarily in the united, latin america, united kingdom, canada and europe, located in the united, throughout the united, south america, asia pacific

Our operating facilities and customers are located primarily in the United States and Canada.

The Company operates 18 manufacturing facilities located throughout the United States, United Kingdom, and Canada.

45) **Short-term credit facilities**: credit facility, revolving credit, senior secured, entered into, secured credit, senior credit, term loan, credit agreement, secured revolving, term debt

The Company has a $150 million secured revolving credit facility.

During 2007, the Company entered into a $20 million senior debt facility agreement.

46) **End-of-year transactions**: ending december, terminates on december, dated december, december we acquired, dated december, investment on december, year ending, company acquired, invested on december, payment was due on december

The outstanding balance at December 31, 2009 and December 31, 2010 was $8.0 million.

On December 15, 2010, the Company renewed the lease agreement effective February 1, 2011.

Continued on next page
For the fiscal years ended September 30, 2010, 2009 and 2008, our loss from joint venture activities, the impairments of our investments in certain of our unconsolidated joint ventures, and the overall equity in loss of unconsolidated joint ventures is as follows:

The table below presents the forecasted amortization expense for intangible assets acquired in all mergers:

On December 8, 2009, the Company acquired VLCY and its subsidiaries.

CSX follows a 52/53 week fiscal reporting calendar which allows every year to consistently end on a Friday.

Stevenson acquired the assets, liabilities and outstanding convertible securities of the Company existing as of September 17, 2010 (the Spin-Off).

As part of the spin off Sunburst and the Company entered into a strategic alliance agreement.

The aircraft consist of 24 A330 aircraft, 20 Boeing 757 aircraft, two Boeing 747-400 aircraft and six A319 aircraft.

The Company is primarily engaged in the sale of aircraft, aircraft parts, leasing of aircraft and engines, and related services.

The Bally Gaming and Systems business unit consists of three separate divisions: Gaming Products, Gaming Operations and Gaming Systems.

Oasis Technologies provides the gaming system and a continuing series of games to deploy in the gaming devices.

We are principally engaged in the exploration development and mining of gold.

The carrying value of the McDonald Gold Project as of December 31, 2004 was $13,359,500.

We are principally engaged in the exploration development and mining of gold.

Marine container lease revenue is expected to decline in the future as the Partnership continues to sell marine containers.

The most significant direct operating costs are wages paid to vessel crews maintenance and repairs and marine insurance.
<table>
<thead>
<tr>
<th><strong>57) Agricultural operations:</strong></th>
<th>crop insurance, crop hail, crop production, agricultural partnerships, crop nutrient, insurance business, crop yields, named peril, crop drying, agricultural market</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reported crop insurance claims are reserved based upon preliminary notice to the Company and investigation of the loss in the field.</td>
<td></td>
</tr>
<tr>
<td>The supply of commodities current pricing and expected new crop quantity and quality affect the timing of the Company’s sales and earnings.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>58) Hotel and lodging operations:</strong></th>
<th>interstate hotels, hotels resorts, service hotels, full service, united states, hotels and resorts, hotel properties, managed hotel, hotels are located, ownership of management</th>
</tr>
</thead>
<tbody>
<tr>
<td>The sales of the hotels in 1997 and 1998 generated gains of $36.0 million.</td>
<td></td>
</tr>
<tr>
<td>The Company renovated and repositioned the hotels as full service upscale unique independent boutique hotels.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>59) Floral products:</strong></th>
<th>crop insurance, crop protection, fresh cut flowers, floral products, specialty retailers, crop nutrient, floral services, named peril, crop year, brooding and weed</th>
</tr>
</thead>
<tbody>
<tr>
<td>A variety of factors affect the supply of flowers in the United States and the price of the Company’s floral products.</td>
<td></td>
</tr>
<tr>
<td>The breadth and diversity of the product line helps to minimize the impact of individual crop earnings fluctuations.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>60) Gaming regulations and violations:</strong></th>
<th>nevada gaming, vegas nevada, gaming authorities, nevada commission, authorities at any time, current stock, examined by the nevada, district court, court for the district, nevada board</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Nevada Commission, the Nevada Board and the various county and municipal licensing agencies are collectively referred to as the Nevada Gaming Authorities.</td>
<td></td>
</tr>
<tr>
<td>The Nevada Gaming Act requires any person who acquires more than 5% of the Company’s voting securities to report the acquisition to the Nevada Commission.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>61) Measurement of natural gas properties:</strong></th>
<th>natural gas properties, natural gas reserves, inherently precise, reserves are inherently, business consists, cost method, full cost, method of accounting, involves a high, estimates of oil and natural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Significant estimates include DDA of proved oil and gas properties.</td>
<td></td>
</tr>
<tr>
<td>Therefore, estimates of natural gas and crude oil reserves are inherently imprecise.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>62) Apparel retail and manufacturing:</strong></th>
<th>women’s apparel, outlet stores, czech republic, apparel group, apparel manufacturers, weekly basis, united states, business segments, dominican republic, men’s apparel</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Retail segment reflects the operations of the Company’s 167 outlet and specialty stores.</td>
<td></td>
</tr>
<tr>
<td>The manufacturers reimburse us weekly monthly or quarterly depending on the manufacturer and the type of program.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>63) Utility operations:</strong></th>
<th>commodity price, square miles, electric service, service to communities, furnishes electric, communities in square, plans in early, companies expect, price risk, miles of western</th>
</tr>
</thead>
<tbody>
<tr>
<td>All of PPL Electrics facilities are located in Pennsylvania.</td>
<td></td>
</tr>
<tr>
<td>The Sponsor serves as the Trusts commodity pool operator and commodity trading advisor.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>64) Gold mining operations:</strong></th>
<th>gold project, gold and silver, gold prices, ounces of gold, gold mine, entered into, gold mineralization, northern territory, gold exploration, gold production</th>
</tr>
</thead>
<tbody>
<tr>
<td>The best of the mineralization showed 4.6 meters of 0.908 gpt gold and 0.24% combined copper and lead.</td>
<td></td>
</tr>
<tr>
<td>During the process of recovering gold, we also produce silver as a byproduct of the gold recovery process.</td>
<td></td>
</tr>
</tbody>
</table>
Appendix A.4  Word Intrusion Tasks

To assess the validity of the topics inferred by the LDA algorithm, we evaluate the semantic meaning and interpretability of the inferred topics using word intrusion tasks as developed by Chang et al. [2009]. We conduct our word intrusion tasks using machine-based and human-subjects procedures. We discuss each procedure in detail below.

A.4.1  Machine-based Intrusion Task

We use a word embedding algorithm to conduct our machine-based intrusion task. Word embedding algorithms are unrelated to LDA and are trained to reconstruct linguistic contexts of words. These algorithms create a vector mapping of the semantics of large text corpora such that words sharing similar contexts are located in close proximity to each other in the vector space, and pairs of words with analogous relationships are located with similar distances between the words.\(^2\) Once the vector mapping is created, the algorithm is able to detect unrelated or intruder words similar to a human subject.

We construct the word intrusion task by taking the 10 most probable words in each of the unaggregated topics based on the word weights. Within each rolling five-year window, we randomly select three of the top 10 words from one topic and one “intruder” word from another randomly selected topic, ignoring words that are in the top ten in both topics. We construct all possible word combinations of this form, resulting in over 12.6 million word sets. We then use the word embedding algorithms to identify the intruder word. We use 25\% as a baseline precision rate to assess task performance given that there are four possible choices in each word set (i.e., a random chance of 1 out of 4 words). The algorithm should correctly identify the unrelated or intruder word at a rate greater than random chance (25\%) if the topic words are coherent and interpretable.

\(^2\)These algorithms use a neural network or statistical algorithm to translate the relative meanings of words into measurable distances. For instance, the algorithms would capture that the relationship between France and Paris is the same as Italy and Rome (country and capital), that Einstein and scientist have the same relationship as Picasso and painter (surname and general occupation), and that the relationship between copper and Cu is the same as Zinc and Zn (element and abbreviation).
Our first run of the task uses the Global Vectors for Word Representation (GloVe) algorithm trained on a general text corpus of websites (Pennington, Socher, and Manning [2014]). This Internet corpus covers 840 billion words from Common Crawl. The corpus is not business specific and thus allows us to examine the coherency of these topics in a general context. We find that the algorithm correctly identifies the intruder word at an average rate of 49.19% across all the word combinations in each rolling window. The precision rate for the rolling windows range from 41.07% to 55.40%. All of the precision rates are significantly higher than random chance (25%) at the 1% level (one-tailed t-test p-value = 0.00 across all windows). These results indicate that the topics inferred by LDA are meaningful in a general context and that the algorithm performs well across time. Figure A.2 presents the annual precision rates for this and subsequent runs of our machine-based intrusion task.

To examine the quality of our topics in a business context, we train a second word embedding algorithm, termed word2vec, on a text corpus of Wall Street Journal (WSJ) articles drawn from ProQuest over the 1999 through 2012 period. The corpus consists of full-text articles randomly selected by day and contains 247.8 million words. To balance model quality and computational difficulty, we retain all words with at least 100 occurrences in the corpus (leaving 243.8 million words). The trained model can be thought of as an expert on general business information. We apply the trained model to the word intrusion task discussed above and constrain the word sets to only those combinations where the model has learned all four words (resulting in a reduced total of 12.3 million word sets). We find that this alternative model accurately identifies 50.0% (p-value = 0.00) of intruder words on average (ranging from 39.7% to 56.8%; p-value = 0.00 across the rolling windows). This result indicates that our LDA topics are also meaningful in a business context and that this generalizes to our entire sample period.

Our final machine-based task provides direct validation of the LDA topics by training

---

3A pretrained GloVe model on this corpus is available at https://nlp.stanford.edu/projects/glove/.  
4The word2vec algorithm, created by Google, computes continuous vector representations of words (Mikolov et al. [2013]).  
5The excluded word sets are those containing words with less than 100 occurrences in the WSJ corpus.
the word2vec algorithm on the entire corpus of text from our sample of 10-K filings. Consequently, the model is trained specifically on the relationships between words and how they are used in 10-K narratives. We test the model on a set of 12.5 million word combinations where the algorithm has learned all four words (we again remove all words with less than 100 occurrences in the 10-K corpus). Our results show that this model performs the best, with an average precision rate of 52.9% \((p\text{-value} = 0.00)\), ranging from 46.37% to 58.7% \((p\text{-value} = 0.00\text{ across all windows})\). The superior performance of this model is not surprising, however, given that the model was trained on the same text corpus from which the LDA topics were derived. Nonetheless, these results provide added support to our inference that the LDA topics are meaningful in the context of our study.\(^6\)

A.4.2 Human-subjects Intrusion Task

Similar to Chang et al. [2009], we use the online labor market of Amazon Mechanical Turk (MTurk) to conduct a human-subject version of our word intrusion task. The MTurk tasks are constructed similarly to our machine-based tasks. We make the procedure practicable by presenting each MTurk participant with a small subset of the word combinations evaluated by the word embedding algorithms. Specifically, once participants access our online experimental materials, they are each presented with 20 word combinations that are selected at random from a pool of 200 combinations. The 200 word sets are randomly selected from the full set of word combinations evaluated by the word embedding algorithms as described above in Section A.4.1. Each participant must then identify the intruder word for each of the 20 word combinations.

Our subject pool consists of 180 MTurk participants who are U.S. citizens, proficient in English, and are at least 18 years of age. We collect demographic information at the end of the task to assess participants familiarity with financial terminology. Our participants are

---

\(^6\)We also find that our precision rates compare well to an objective word embedding task of matching countries with country capitals using both the GloVe algorithm and the WSJ-trained word2vec algorithm (average precision rates of 57% and 58%, respectively).
34 years old on average, with roughly 85% completing some college and above. Sixty-four percent of the participants report using financial statements at least once to evaluate a firm’s performance. Lastly, 60% of the subject pool have bought or sold a firm’s common stock or debt securities, while 67% plan on investing in a firm’s securities within the next five years. Overall, these demographics suggest that our MTurk participants have reasonable levels of financial knowledge, consistent with the evidence in Farrell, Grenier, and Leiby [2017].

The results from the human-subjects task indicate that participants correctly identify the intruder word about 40% of the time. This precision rate is statistically higher than random chance at the 1% level (one-tailed $t$-test $p$-value = 0.00). Although the human precision rate is lower than that achieved by the machine algorithm, it still compares well when we consider that the word embedding algorithms were trained on entire corpora of Internet, WSJ, and 10-K narratives. We acknowledge, however, that the lower rate could reflect noise in the topics derived from the LDA model.
This graph illustrates the annual precision rates of the machine-based word intrusion tasks discussed in Appendix A.4.1. The annual rates are based on the unaggregated topics discovered by the LDA algorithm over the five-year window immediately preceding each year. The word intrusion tasks based on the general Internet corpus (Common Crawl) were performed using the GloVe word embedding algorithm, while the intrusion tasks based on the Wall Street Journal (WSJ) and 10-K filing corpora were performed using the word2vec algorithm.
Appendix A.5  Logit Regression Convergence

We implement a seven-step procedure to ensure proper convergence of our logistic regressions and to mitigate problems from completeness and quasi-completeness. The seven steps are as follows:

1. Use a QR decomposition of our matrix of independent variables to determine an order of columns which are most likely to be degenerate.

2. If the number of independent variables is greater than the number of cases in which the dependent variable (misreport) is equal to 1, then remove independent variables in the order determined in step 1 until the number of remaining independent variables is equal to the number of cases in which the dependent variable is equal to 1.

3. Run a 50 iteration logit using a Newton-Raphson solver.

4. Check for convergence and make sure there were no signs of quasi-completeness (massively inflated standard errors), if failed, then go to step 7.

5. If step 4 is passed, then run a 500 iteration logit using a Newton-Raphson solver with the parameters from step 3 as an initial parameter set.

6. Ensure the regression in step 5 converges. If the regression converges, then the operation is successful. If convergence fails, then go to step 7.

7. Remove 1 independent variable following the order from step 1, go back to step 2.
Appendix A.6  Robustness Tests

A.6.1  Frequency of Misreporting Events Over Time

The frequency of AAERs declines substantially after 2005 (see Table 1 of the paper), due to the length of time necessary for the SEC to investigate and process an enforcement action and to possible changes in the SEC’s enforcement priorities after the 2008 financial crisis (Bao et al. [2019]). The frequency of irregularities in the AA sample also declines from 2005 onward following an uptick in restatements after enactment of the 2002 Sarbanes-Oxley Act (SOX). To ensure that our results reflect a complete set of true misreporting events, we limit the AAER (AA) sample to firm-years from 1994 to 2005 (2005 to 2012). This restriction narrows our set of prediction years to 1999 to 2005 (2010 to 2012) in the AAER (AA) sample. We then replicate our detection models using this smaller set of prediction years. In untabulated results, we continue to find that topic provides significant incremental value in detecting misreporting in both samples. Consistent with the results reported in the paper, the topic measure improves prediction accuracy substantially when added to both the stand-alone F-score (Style) model and the joint F-score and Style model in the AAER (AA) sample.

A.6.2  Technical Filing Amendments

We conduct further screens of the 10-K/A irregularity sample to identify filing amendments that might be more technical or routine in nature. The most common type of technical amendments are revisions to proxy statement information disclosed in Part III (Items 10 – 14) of Form 10-K (Ntiamoah and Plante [2019]). We screen for proxy amendments by searching the text of the explanatory note in the amended filing for phrases such as “proxy statement”, “part III”, and “item 1#”. We then remove any irregularity event from the 10-K/A sample that refers to a proxy-related amendment based on our search string. This process reduces the percentage of irregularities to 1.30% of the full 10-K/A sample; a percentage rate that
is closer to the misreporting rate found in the AAER sample (see Table 1 of the paper). The classification performance of our models is slightly weaker for this reduced set of 10-K/A irregularities. But despite this, we continue to find evidence that topic adds significant incremental value to detecting financial misreporting. For instance, the addition of topic to the benchmark F-score and Style model increases classification accuracy for high-risk observations by 8% and 30% at the 95th and 99th percentile cut-offs, respectively. These rates of relative improvement are higher than the rates reported in Section 4.3 of the paper and likely reflect the use of a less noisy 10-K/A sample.

A.6.3 Unintentional Errors

We investigate our models’ ability to detect misreporting events involving unintentional errors (i.e., misreporting stemming from accounting mistakes, data errors, and general presentation or disclosure errors). This analysis serves partly as a validation test of our topic measure. Since errors do not reflect explicit intent to report misleading information, we expect disclosure content to vary less around these events, thus reducing the incremental predictive value of topic.

To conduct this analysis, we follow prior research and rely on the AA restatement database given its clear classification of misreporting events due to accounting errors. We classify a restatement as an unintentional error if the event is coded in the AA database as a clerical error (res_clerical_errors = 1) and not also flagged as a financial fraud, irregularity, or misrepresentation (res_fraud = 0) or as an “other significant issue” (res_other = 0). We allow the sample to overlap with restatements arising from GAAP application failures and those involving regulatory investigations since clerical errors can still arise from or be identified through these mechanisms.

Consistent with our conjecture, we find that the incremental value of topic is economically

---

7Our screening process leads to a conservative sample as it unavoidably excludes observations with a material revision or omission elsewhere in the 10-K filing. A manual check of 20 random observations that are screened out identifies 4 cases with material revisions to other parts of the filing stemming from reporting problems that fit the classification criteria in Hennes, Leone, and Miller [2008]
smaller relative to our reported results, though models that include topic still perform better in detecting errors. Thus, given that some errors can escalate to intentional misreporting (perhaps to conceal the error), it seems that topic provides some added value as an early warning sign.

A.6.4 Using MD&A Text

Several text-based studies of financial misreporting focus primarily on the MD&A section of the 10-K filing (see e.g., Cecchini et al. [2010a], Hoberg and Lewis [2017], Purda and Skillicorn [2015]). We therefore investigate whether our results differ when we restrict our LDA and textual analysis to the MD&A section. We reconstruct our topic and Style variables using only the text extracted from the MD&A section (see Appendix A.1 of this appendix for further details). The incremental value of topic in the AAER sample is weaker compared to the results reported in the paper. However, topic continues to provide significant incremental power for detecting misreporting in the AA and 10-K/A irregularity samples. These results reaffirm prior arguments in the literature that 10-K narratives in their entirety provide additional content that is useful for drawing inferences in text-based research (see Li [2010b], Loughran and McDonald [2016]).

A.6.5 Regularized Logit Regression

Due to the relative rarity of misreporting events, there is some concern of overfitting given the large number of predictors in our models (see Perols et al. [2017]). We address this issue by using an L1 regularized logistic regression to re-estimate the results. The L1 regularization approach applies a penalty for increasing the number of independent variables, thereby controlling for biases arising from model overfitting. All of our out-of-sample results provide inferences that are similar to the results reported in the paper.
A.6.6 Additional Financial and Textual Style Measures

We test whether our inferences are robust to the inclusion of other financial and textual style measures used to detect financial misreporting in prior studies. The additional financial measures include a standard set of financial ratios, bankruptcy prediction measures, and the variables used to estimate the Beneish M-Score (Beneish [1997,1999], Cecchini et al. [2010a,b]). The expanded set of textual style features includes alternative measures of readability, length, repetition, lexical variety, and deviation from the Benford distribution (Amiram, Bozanic, and Rouen [2015], Goel et al. [2010], Goel and Gangolly [2012]). The replicated results show no change in the tenor of the results reported in the paper.

A.6.7 Raw topic Measure

Our final sensitivity check uses the raw topic proportions instead of the normalized proportions described in the paper. This alternative measurement approach increases the variance of topic, as the measure is now influenced by the amount of text in the document. The prediction results across all three misreporting samples are qualitatively similar to the results reported in the paper. The raw topic measure performs slightly weaker in our comparative tests, but the incremental predictability of topic remains strong in general. Thus, a raw measure of disclosure content, rather than the proportion, is also useful for detecting financial misreporting.
References


