

Stylistic Features in Corporate Disclosures and their Predictive Power

Philipp Heinrich

Chair of Computational Corpus Linguistics

Friedrich-Alexander University of Erlangen-Nuremberg

APCLC 2018 – September 17, 2018



Introduction



Introduction

- object of analysis: annual reports submitted to the U.S. SEC's EDGAR filing system ("**form 10-k**")
- **corpus of 76,278 documents** from between January 2006 and December 2015
- partly structured information (XBRL data), actual textual content: without semantic mark-up (split into up to 20 items)
- related work: influence on stock market (item 7: MDA)
- here: calculation of **quantitative linguistic features** (sentiment, readability, Biber's stylistic features) of (almost) all items
- long-term goal: facilitate interpretation of text type; "read between the lines", get "hidden" information

Table of Contents

Introduction

Corpus and Methodology

- Items

- Preprocessing

- Quantitative Linguistic Features

Explorative Analysis

- Descriptive Figures

- Clustering

Conclusion

Corpus and Methodology



**UNITED STATES
SECURITIES AND EXCHANGE COMMISSION**
Washington, D.C. 20549

FORM 10-K

(Mark One)

ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934

For the fiscal year ended December 31, 2016

or

TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934

For the transition period from _____ to _____

Commission File No. 000-22513

AMAZON.COM, INC.

(Exact name of registrant as specified in its charter)

Delaware
(State or other jurisdiction of
incorporation or organization)

91-1646860
(I.R.S. Employer
Identification No.)

410 Terry Avenue North
Seattle, Washington 98109-5210
(206) 266-1000

(Address and telephone number, including area code, of registrant's principal executive offices)

Securities registered pursuant to Section 12(b) of the Act:

Title of Each Class
Common Stock, par value \$ 0.01 per share

Name of Each Exchange on Which Registered
NASDAQ Global Select Market

Securities registered pursuant to Section 12(g) of the Act:
None

Indicate by check mark if the registrant is a well-known seasoned issuer, as defined in Rule 405 of the Securities Act. Yes No

Indicate by check mark if the registrant is not required to file reports pursuant to Section 13 or Section 15(d) of the Exchange Act. Yes No

Indicate by check mark whether the registrant (1) has filed all reports required to be filed by Section 13 or 15(d) of the Securities Exchange Act of 1934 during the preceding 12 months (or for such shorter period that the registrant was required to file such reports), and (2) has been subject to such filing requirements for the past 90 days. Yes No

Indicate by check mark whether the registrant has submitted electronically and posted on its corporate Web site, if any, every Interactive Data File required to be submitted and posted pursuant to Rule 405 of Regulation S-T during the preceding 12 months (or for such shorter period that the registrant was required to submit and post such files). Yes No

Indicate by check mark if disclosure of delinquent filers pursuant to Item 405 of Regulation S-K is not contained herein, and will not be contained, to the best of registrant's knowledge, in definitive proxy or information statements incorporated by reference in Part III of this Form 10-K or any amendment to this Form 10-K.

Indicate by check mark whether the registrant is a large accelerated filer, an accelerated filer, a non-accelerated filer, or a smaller reporting company. See definitions of "large accelerated filer," "accelerated filer" and "smaller reporting company" in Rule 12b-2 of the Exchange Act.

Large accelerated filer

Accelerated filer

Non-accelerated filer

(Do not check if a smaller reporting company)

Smaller reporting company

Indicate by check mark whether the registrant is a shell company (as defined in Rule 12b-2 of the Exchange Act). Yes No

Aggregate market value of voting stock held by non-affiliates of the registrant as of June 30, 2016

\$

280,129,378,534

Number of shares of common stock outstanding as of January 25, 2017

477,170,618

DOCUMENTS INCORPORATED BY REFERENCE

The information required by Part III of this Report, to the extent not set forth herein, is incorporated herein by reference from the registrant's definitive proxy statement relating to the Annual Meeting of Shareholders to be held in 2017, which definitive proxy statement shall be filed with the Securities and Exchange Commission within 120 days after the end of the fiscal year to which this Report relates.

Items in the 10k-forms

Part I

Item 1 Business

Item 1A Risk Factors

Item 1B Unresolved Staff Comments

Item 2 Properties

Item 3 Legal Proceedings

Item 4 Mine Safety Disclosures

Items in the 10k-forms

Part II

- Item 5** Market for Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases
- Item 6** Selected Financial Data
- Item 7** Management's Discussion and Analysis of Financial Condition and Results of Operations
- Item 7A** Quantitative and Qualitative Disclosures About Market Risk
- Item 8** Financial Statements and Supplementary Data
- Item 9** Changes in and Disagreements With Accountants on Accounting and Financial Disclosure
- Item 9A** Controls and Procedures
- Item 9B** Other Information

Items in the 10k-forms

Part III

Item 10 Directors, Executive Officers and Corporate Governance

Item 11 Executive Compensation

Item 12 Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters

Item 13 Certain Relationships and Related Transactions, and Director Independence

Item 14 Principal Accounting Fees and Services

Part IV

Item 15 Exhibits, Financial Statement Schedules

Item 16 Form 10-K Summary

Preprocessing

1. extract HTML part (easy)
2. extract items (more difficult than expected)

HTML source code snippet

```
<TABLE width="100%" border="0" cellpadding="0" cellspacing="0"
style="font-size: 10pt; font-family: Arial, Helvetica;
color: #000000; background: #FFFFFF">
<TR>
  <TD width="8%"></TD>
  <TD width="92%"></TD>
</TR>
<TR valign="top">
  <TD>
    <B><FONT style="font-family: 'Times New Roman',
Times">Item&#160;2.&#160;&#160;&#160;&#160;</FONT></B>
  </TD>
  <TD>
    <A name='132'></A><B><FONT style="font-family:
'Times New Roman', Times">Properties</FONT></B>
  </TD>
</TR>
</TABLE>
```

Preprocessing

1. extract HTML part (easy)
2. extract items (more difficult than expected)
 - some more or less well-maintained code available for specific items, e. g. <https://github.com/iammrhelo/edgar-10k-mda>
 - other research teams deal with the same issues, e. g. University of Ljubljana (government-funded project “Influence of formal and informal corporate communications on capital markets”)
 - our item-extractor:
 - open-source Python code
 - cascade of regular expressions
 - manually created gold-standard
 - not yet available ☹️

Preprocessing

1. extract HTML part (easy)
2. extract items (more difficult than expected)
 - some more or less well-maintained code available for specific items, e. g. <https://github.com/iammrhelo/edgar-10k-mda>
 - other research teams deal with the same issues, e. g. University of Ljubljana (government-funded project “Influence of formal and informal corporate communications on capital markets”)
 - our item-extractor:
 - open-source Python code
 - cascade of regular expressions
 - manually created gold-standard
 - not yet available 😞

Corpus composition – items

item	total number	filtered
1	63,577	56,977
1A	53,553	47,953
2	21,224	18,686
3	20,827	18,619
5	51,964	46,450
7	57,863	51,789
7A	31,069	27,693
8	33,513	30,009
9A	57,265	51,346
10	33,885	30,334

Table: Number of observations in the corpus for all items with more than 20,000 admissible observations; total absolute frequencies are given in the first column, frequencies in the category-filtered corpus in the second column.

Corpus composition – filing companies' industry classification

industry	total number
<i>finance</i>	21,914
<i>manufacturing</i>	23,199
<i>mining</i>	5,856
<i>services</i>	12,233
<i>TCEGS</i>	5,562
<i>other</i>	6,635
<i>unknown</i>	879
total	76,278

Table: Distribution of industry categories of the filing companies. *TCEGS* is short for *Transportation, Communications, Electric, Gas, and Sanitary Services*.

Sentiment & Subjectivity

- SentiKLUE (Evert et al., 2014)
- TextBlob (<https://textblob.readthedocs.io/en/dev/>)

Readability

- Fog Index (Gunning, 1952):

$$l_{\text{fog}} = 0.4 \left(n_w/n_s + 100 \cdot n_c/n_w \right)$$

- Flesh-Kincaid grade-level (Kincaid, 1975):

$$l_{\text{fk}} = 0.39 \left(n_w/n_s \right) + 11.8 \left(n_y/n_w \right) - 15.59$$

- frequently discussed (Bonsall et al., 2017; Loughran and McDonald, 2014, 2016; Si and Callan, 2001)

Sentiment & Subjectivity

- SentiKLUE (Evert et al., 2014)
- TextBlob (<https://textblob.readthedocs.io/en/dev/>)

Readability

- Fog Index (Gunning, 1952):

$$l_{\text{fog}} = 0.4 \left(n_w/n_s + 100 \cdot n_c/n_w \right)$$

- Flesh-Kincaid grade-level (Kincaid, 1975):

$$l_{\text{fk}} = 0.39 \left(n_w/n_s \right) + 11.8 \left(n_y/n_w \right) - 15.59$$

- frequently discussed (Bonsall et al., 2017; Loughran and McDonald, 2014, 2016; Si and Callan, 2001)

Sentiment & Subjectivity

- SentiKLUE (Evert et al., 2014)
- TextBlob (<https://textblob.readthedocs.io/en/dev/>)

Readability

- Fog Index (Gunning, 1952):

$$l_{\text{fog}} = 0.4 \left(n_w/n_s + 100 \cdot n_c/n_w \right)$$

- Flesh-Kincaid grade-level (Kincaid, 1975):

$$l_{\text{fk}} = 0.39 \left(n_w/n_s \right) + 11.8 \left(n_y/n_w \right) - 15.59$$

- frequently discussed (Bonsall et al., 2017; Loughran and McDonald, 2014, 2016; Si and Callan, 2001)

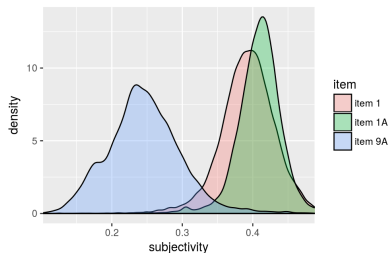
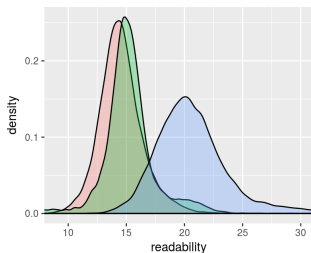
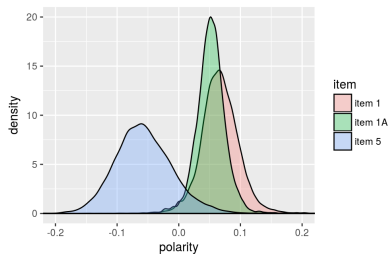
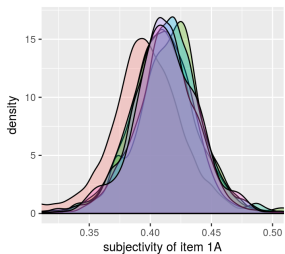
Biber's Stylistic Features

- “Variation across Speech and Writing” (Biber, 1988), “The Multi-dimensional Approach to Linguistic Analyses”
- dimensions:
 1. the opposition between **involved** and **informational** discourse
 2. the opposition between **narrative** and **non-narrative** concerns
 3. the opposition between **context-independent** and **context-dependent** discourse
 4. overt expression of **persuasion**
 5. opposition between **abstract** and **non-abstract** information
 6. **on-line informational elaboration**
- MAT implementation by Andrea Nini (Nini, 2015):
<https://sites.google.com/site/multidimensionaltagger/>
- grammatical annotation (Stanford Tagger + minor modifications)

Explorative Analysis



Descriptive Figures: Sentiment & Readability



Descriptive Figures: Sentiment & Readability

- statistical procedure: ANOVA + Tukey HSD
- difference in means of sentiment and polarity significant across industry categories:
 - financial companies: most objective, most complex language
- difference in means of sentiment and polarity significant across items:
 - item 1A (description of risk factors): most subjective
 - item 5 (explanation of highs and lows of the company's stock): only item with negative sentiment polarity
 - item 9A (conclusions of company's principal officers): most objective, most complicated

Descriptive Figures: Biber

dimension	1	2	3	4	5	6
avg. score	-10.5	-4.2	9.0	-4.3	0.7	-1.7

Table: Average scores across all items and industries in Biber's six stylistic dimensions.

- high values in absolute terms:
 - informationally dense (dimension 1)
 - non-narrative (dimension 2)
 - context-independent (dimension 3)
 - no overt expressions of persuasion (dimension 4)
- mining companies: on average higher scores in all dimensions
- item 7A (quantitative information about market risk) and item 9A (conclusions of company's principal officers): especially low in dimension 1
- item 3 (legal proceedings): especially high in dimension 4

Clustering: Methodology

- goal: show that the quantitative linguistic features have predictive power
- clustering rather than statistical categorization
- representation of each document by a 90-dimensional vector (10 items times 9 scores)
- t-distributed stochastic neighbour embeddings: projection onto a two-dimensional plane
- only use documents with only admissible items

industry	total	t-SNE input
<i>finance</i>	21,914	66
<i>manufacturing</i>	23,199	265
<i>mining</i>	5,856	38
<i>services</i>	12,233	67
<i>TCEGS</i>	5,562	53
<i>total</i>		489

Clustering: Methodology

- goal: show that the quantitative linguistic features have predictive power
- clustering rather than statistical categorization
- representation of each document by a 90-dimensional vector (10 items times 9 scores)
- t-distributed stochastic neighbour embeddings: projection onto a two-dimensional plane
- only use documents with only admissible items

industry	total	t-SNE input
<i>finance</i>	21,914	66
<i>manufacturing</i>	23,199	265
<i>mining</i>	5,856	38
<i>services</i>	12,233	67
<i>TCEGS</i>	5,562	53
<i>total</i>		489

t-SNE projection of quantitative linguistic features

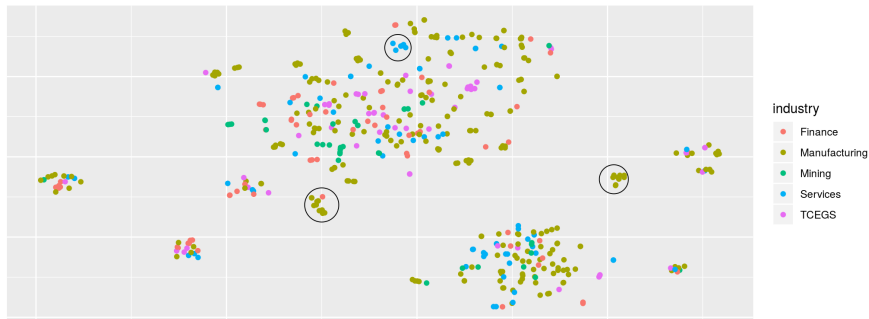


Figure: t-SNE projection of quantitative linguistic features (sentiment, polarity, readability, and the Biber features) of 489 documents onto a 2-dimensional plane. The categories of the filing companies are highlighted by colour, and several meaningful clusters can be identified.



manufacturing cluster slightly left of the middle of the figure.

t-SNE projection of quantitative linguistic features

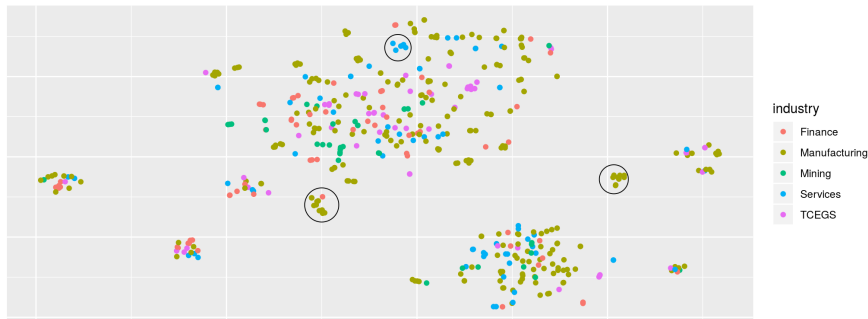



Figure: t-SNE projection of quantitative linguistic features (sentiment, polarity, readability, and the Biber features) of 489 documents onto a 2-dimensional plane. The categories of the filing companies are highlighted by colour, and several meaningful clusters can be identified.

 *manufacturing* cluster slightly left of the middle of the figure.

A look into the corpus

- *KIMBERLY CLARK CORP* (CIK 55785) on 22 February, 2008:

Kimberly-Clark Corporation was incorporated in Delaware in 1928. The Corporation is a global health and hygiene company focused on product innovation and building its personal care, consumer tissue, K-C Professional & Other and health care operations. The Corporation is principally engaged in the manufacturing and marketing of a wide range of health and hygiene products around the world.

A look into the corpus

- *Revett Minerals Inc.* (CIK 1404592) on March 28, 2012:

Revett Minerals Inc. (“Revett Minerals”) was incorporated under the Canada Business Corporations Act in August 2004 to acquire Revett Silver Company (“Revett Silver”), a Montana corporation, and undertake a public offering of its common stock in Canada, transactions that were completed in February 2005.

- *EASTERN CO* (CIK 31107) on March 13, 2015:

The Eastern Company (the “Company”) was incorporated under the laws of the State of Connecticut in October, 1912, succeeding a co-partnership established in October, 1858. The business of the Company is the manufacture and sale of industrial hardware, security products and metal products from six U.S. operations and seven wholly-owned foreign subsidiaries. The Company maintains thirteen physical locations.

Conclusion



Conclusion

- processing of 10-k forms: item extraction
 - ☞ open source version available soon
- calculation of quantitative linguistic features
 - ☞ proof of concept
- clustering experiment
 - ☞ quantitative linguistic features have predictive power

Future Research

- improve sentiment & readability measures
- calculate further quantitative features, e. g. uncertainty
- link quantitative linguistic data with quantitative financial data (XBRL + external databases)

Conclusion

- processing of 10-k forms: item extraction
 - ☞ open source version available soon
- calculation of quantitative linguistic features
 - ☞ proof of concept
- clustering experiment
 - ☞ quantitative linguistic features have predictive power

Future Research

- improve sentiment & readability measures
- calculate further quantitative features, e. g. uncertainty
- link quantitative linguistic data with quantitative financial data (XBRL + external databases)

Thanks for listening.
Questions?

References



- Douglas Biber. *Variation across Speech and Writing*. Cambridge University Press, Cambridge, 1988.
- Samuel B. Bonsall, Andrew J. Leone, Brian P. Miller, and Kristina Rennekamp. A plain English measure of financial reporting readability. *Journal of Accounting and Economics*, 63(2-3):329–357, April 2017.
- Stefan Evert, Thomas Proisl, Paul Greiner, and Besim Kabashi. SentiKLUE: Updating a polarity classifier in 48 hours. In Preslav Nakov and Torsten Zesch, editors, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 551–555, Dublin, 2014. Association for Computational Linguistics.
- R. Gunning. *The Technique of Clear Writing*. McGraw-Hill International Book Co. McGraw-Hill International Book Co., New York, NY, 1952.
- J.P. Kincaid. *Derivation of New Readability Formulas: (automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Research Branch report. Chief of Naval Technical Training, Naval Air Station Memphis, 1975.
- Tim Loughran and Bill McDonald. Measuring Readability in Financial Disclosures. *The Journal of Finance*, 69(4): 1643–1671, 2014.
- Tim Loughran and Bill McDonald. Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, 54(4):1187–1230, September 2016.
- Andrea Nini. Multidimensional Analysis Tagger (version 1.3)., 2015.
- Luo Si and Jamie Callan. A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management, CIKM '01*, pages 574–576, New York, NY, USA, 2001. ACM.