A Naïve Bayesian Machine Learning Approach to Corporate Filings

COMP 488 - Machine Learning

Jose Luis Rodriguez

Department of Computer Science
Loyola University Chicago

September 4, 2017

The Journal of Accounting Research

- The Journal of Accounting Research has been published since 1963 by the Accounting Research Center at the University of Chicago Booth School of Business
- Is a general-interest accounting journal. It publishes original research in all areas of accounting that utilizes tools from basic disciplines such as economics, statistics, psychology, and sociology
- Research typically uses analytical, empirical archival, experimental, and field study methods and addresses economic questions
- Areas: Accounting, auditing, taxation, and related fields such as corporate finance, investments, capital markets, law, and information economics

Security and Exchange Commission[1]

- The mission of the SEC is to protect investors; maintain fair, orderly, and efficient markets; and facilitate capital formation. The SEC strives to promote a market environment that is worthy of the public's trust
- in 1980, the SEC mandated that public companies include in their annual reports a section for Management's Discussion and Analysis of Financial Condition and Results of Operations (MD&A)
- The MD&A is intended to assess an enterprise's liquidity, capital resources, and operations in a way that many investors can understand

[1] Source: http://www.jstor.org/stable/40929537

Security and Exchange Commission[2]

- One of the SEC's goals in mandating the MD&A was to make public the information about predictable future events and trends that may affect future operations of the business
- The safe harbor provisions of the Private Securities Litigation Reform Act of 1995 encourage more forward-looking information and should make Management Discussion and Analysis disclosures more informative
- The MD&A might not be as informative as intended for several reasons. Companies have concerns over proprietary costs and uncertainties about the judicial interpretation of safe harbor protection

[2] Source: http://www.jstor.org/stable/40929537

Management Discussion and Analysis (MD&A)[3]

- The Management Discussion and Analysis (MD&A) is the section of a company's annual report in which management provides an overview of the previous year?s operations and how the company performed financially
- Management also discusses the upcoming year by outlining future goals and approaches to new projects.
- The MD&A is an important document for analysts and investors who want to review the company's financial fundamentals and management performance
- Annual reports (10-K), quarterly report and (10-Q) filings

[3] Source: http://www.investopedia.com/terms/m/mdanalysis.asp

Forward-Looking Statements (FORM 10-K)[4]

This Annual Report on Form 10-K contains forward-looking statements. When used in this Annual Report on Form 10-K, the words "may," "could," "estimate," "intend," "continue," "believe," "expect" or "anticipate" and similar expressions identify forward-looking statements. Although we believe that our plans, intentions, and expectations reflected in any forward-looking statements are reasonable, these plans, intentions, or expectations may not be achieved ...

[4] Source: http://www.sec.gov/cgi-bin/browse-edgar

Item 7. Management's Discussion and Analysis of Financial Condition and Results of Operations.

Overview of Current Operations

We were incorporated on May 8, 2015 as Infinity Distribution, Inc., a Nevada corporation. We consider ourselves to be an emerging growth company under applicable federal securities laws and will be subject to reduced public company reporting requirements. The Company is planning to import and export furniture, cacoa and home goods.

Results of Operations

For the fiscal years ended May 31, 2017 and May 31, 2016, the Company recognized no revenues. For the fiscal year ended May 31, 2017, the Company incurred total operating expenses of $200,399, which consists of $395 in depreciation, $153,750 in executive compensation, $15,596 in general and administrative expenses, and $30,658 in professional fees. This compares to the fiscal year ended May 31, 2016, where the Company incurred total operating expenses of $227,723, which consisted of $394 in depreciation, $156,500 in executive compensation, $24,476 in general and administrative expenses, and $46,353 in professional fees.

...

Going Concern

Our ability to continue as a going concern is contingent upon the successful completion of additional financing arrangements and our ability to achieve and maintain profitable operations. Therefore, management plans to raise equity capital to finance the operating and capital requirements of the Company. While the Company is devoting its best efforts to achieve the above plans, there is no assurance that any such activity will generate funds that will be available for operations. These conditions raise substantial doubt about the Company's ability to continue as a going concern. Summary of any product research and development that we will perform for the term of our plan of operation.

...

Significant changes in the number of employees

As of August 28, 2017, we have two employees who also serve as our officers. We are dependent upon our officers for our future business development. As our operations expand we anticipate the need to hire additional employees, consultants and professionals; however, the exact number is not quantifiable at this time.

...

The Research & Abstract

The Information Content of Forward-Looking Statements in Corporate Filings A Naive Bayesian Machine Learning Approach

This paper examines the information content of the forward-looking statements in the Management Discussion and Analysis section (MD&A) of 10-K and 10-Q filings using a Naive Bayesian machine learning algorithm

Abstract

- The average tone of the forward-looking statements is positively associated with future earnings even after controlling for other determinants of future performance
- Despite increased regulations aimed at strengthening MD&A disclosures, there is no systematic change in the information content of MD&As over time

Abstract

- Firms with better current performance, lower accruals, smaller size, lower market-to-book ratio, less return volatility, lower MD&A Fog index, and longer history tend to have more positive forward-looking statements
- The tone measures based on three commonly used dictionaries (Diction, General Inquirer, and the Linguistic Inquiry and Word Count) do not positively predict future performance
- This result suggests that these dictionaries might not work well for analyzing corporate filings

Text Classification: Dictionary Approach

The dictionary approach uses a "mapping" algorithm in which a computer program reads the text and classifies words (or phrases) into different categories based on predefined rules (i.e., dictionary – Diction, General Inquirer, and the Linguistic Inquiry and Word Count).

Text Classification: Naïve Bayesian

Pioneered by computer scientists and mathematicians, relies on statistical techniques to infer the content of text and classify documents based on statistical inference. For instance, the algorithm may calculate the statistical correlation between the frequency of some keywords and the document type to draw inferences.

Naïve Bayesian Algorithm

- A given sentence is first reduced to a list of words (words) with each word weighted in some fashion (e.g., by frequency in the sentence)
- The goal is to classify the sentence into a specific category (cat) from a set of all possible categories (cats)
- In this research there are four possible tone categories: positive, negative, neutral, and uncertain

Accounting Meets Machine Learning

The Naïve Bayesian algorithm chooses the best category by solving the following problem:

$$cat^* = \underset{cat \in cats}{\text{argmax}} \frac{P(words|cat)P(cat)}{P(words)} \tag{1}$$

Accounting Meets Machine Learning

Formula used in the document categorization algorithm on this research:

$$cat* = \underset{cat \in cats}{\text{argmax}}\, P(w_1|cat) * P(w_2|cat) * \cdots * P(w_n|cat) * P(cat) \quad (2)$$

- Since $P(words)$ does not change over the range of categories
- If $w_1, w_2, \ldots, w_n$ are the words in the document and their probability of appearing in a sentence is assumed to be independent

TABLE 4

Correlation Matrix (p-values in Parentheses)

Variables	TONE	PROFIT_ TONE	LIQUIDITY_ TONE	OTHER_ TONE	PROFIT_ PCT	LIQUIDITY_ PCT	EARN	RET	ACC	SIZE	MTB	RETVOL	FOG	FIRMAGE
TONE	1.00													
PROFIT_TONE	0.78	1.00												
	(0.00)													
LIQUIDITY_TONE	0.54	0.29	1.00											
	(0.00)	(0.00)												
OTHER_TONE	0.49	0.26	0.23	1.00										
	(0.00)	(0.00)	(0.00)											
PROFIT_CAT	−0.47	−0.21	0.03	−0.14	1.00									
	(0.00)	(0.00)	(0.00)	(0.00)										
LIQUIDITY_CAT	0.53	0.23	0.01	0.22	−0.83	1.00								
	(0.00)	(0.00)	(0.05)	(0.00)	(0.00)									
OTHER_CAT	−0.05	−0.02	−0.06	−0.12	−0.36	−0.21	1.00							
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)								
EARN	0.15	0.11	0.16	0.07	−0.03	0.02	0.02	1.00						
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)							
RET	−0.01	−0.01	0.00	0.00	−0.01	−0.00	0.02	0.08	1.00					
	(0.06)	(0.00)	(0.58)	(0.68)	(0.00)	(0.87)	(0.00)	(0.00)						
ACC	0.05	0.03	0.06	0.01	−0.01	0.02	−0.00	0.75	0.05	1.00				
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.16)	(0.00)	(0.00)					
SIZE	0.05	0.07	−0.01	−0.01	−0.05	−0.02	0.12	0.22	0.10	0.09	1.00			
	(0.00)	(0.00)	(0.00)	(0.03)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)				
MTB	−0.17	−0.13	−0.12	−0.08	0.10	−0.09	−0.02	−0.21	0.22	−0.06	0.15	1.00		
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)			
RETVOL	−0.26	−0.20	−0.17	−0.09	0.16	−0.11	−0.10	−0.25	0.17	−0.11	−0.33	0.20	1.00	
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)		

(*Continued*)

This table show the pair-wise Pearson correlation coefficients of selected variables with the p-values testing whether the correlation coefficients are significantly different from 0.

Table 4. Correlation Matrix

The variables are defined as follows:

- TONE is the average tone of FLS of a firm-quarter. A forward-looking sentence's tone has a value of 1 if the learning algorithm classifies the sentence as positive, 0 if neutral, and -1 if negative or uncertain
- PROFIT_TONE is the average tone of the FLS of a firm-quarter that are about profits or operations (i.e the statements that are classified as Categories 1 to 4 as defined in appendix C)
- ...

Conclusions

- This paper examines the implications of the FLS in the MD&A section of 10-Q and 10-K filings for future performance
- A Naïve Bayesian machine learning algorithm was used to categorize the tone and content of FLS from more than 140,000 10-Q and 10-K filings between 1994 and 2007
- The tone of the FLS is a function of current performance, accruals, firm size, MTB ratio, return volatility, MD&A Fog, and firm age

Conclusions

- The tone of the FLS is positively correlated with future performance and has explanatory power incremental to other variables
- The informativeness of MD&As has not changed systematically over time despite continuous efforts from the SEC to strengthen MD&A disclosures
- When managers warn in the MD&A about the future performance implications of accruals, accruals are less likely to be mispriced by investors

Conclusions

- MD&A tone measures based on dictionary approaches do not associate positively with future performance
- The Bayesian tone measure remains positively and significant associated with future earnings even when the dictionary-based tone measures are controlled
- The dictionary approach might not work well for analyzing the tone of corporate filings

Appendix

APPENDIX A2

Sample Textual Analysis Papers

Paper	Text Analyzed	Name of Method	Method	Firm-Level Measure
This paper	Forward-looking MD&A	Naïve Bayes	Classify sentences as positive/negative/ neutral/uncertain	Average of sentence tones
Davis, Piger, and Sedor [2005]	Press release	DICTION	Classify words as optimistic/pessimistic	% of words
Kothari, Li, and Short [2009]	MD&A, analyst reports, etc.	General Inquirer	Classify words as optimistic/pessimistic	% of words
Li [2008]	10-Ks and different sections	Fog/word count and LIWC	Classify at word level and averaged across sentences	% of words and average length of sentence
Henry [2008]	Earnings release	Customized dictionary	Classify words as positive/negative	% of words
Matsumoto, Pronk, and Roelofsen [2008]	Conference calls	Customized dictionary and LIWC	Classify words into forward-looking	% of words
Tetlock, Saar-Tsechansky, and Macskassy [2007]	News articles	General Inquirer	Classify words as optimistic/pessimistic	% of words
Feldman et al. [2009]	MD&A	Customized dictionary	Classify words as positive/negative	% of words
Rogers, Buskirk, and Zechman [2009]	Earnings announcements	DICTION	Classify words as optimistic/pessimistic	% of words
Henry and Leone [2010]	Earnings releases	Customized dictionary	Classify words as positive/negative	% of words

APPENDIX C
Classification Categories

- Category 1: Sales/revenues/market condition/market position/consumer demand/competition/pricing/new contract
- Category 2: Cost/expense/reserves for contingent liability/asset impairment/goodwill impairment
- Category 3: Profit/income/performance results/margin
- Category 4: Operations/productions/general business
- Category 5: Liquidity: interest coverage/cash balance/working capital conditions
- Category 6: Investment—general capital expenditure; M&A/divestiture/discontinued operation
- Category 7: Financing—debt/equity/dividend/repurchase
- Category 8: Litigation/lawsuit
- Category 9: Employee relations/retention/hiring/union relations
- Category 10: Regulations (e.g., environment laws)/income tax/government relation
- Category 11: Accounting method/accounting estimation assumptions/auditing/internal control
- Category 12: Other: Boilerplate/legal statement/standard statement