Predictability of Textual Financial Reports on Corporate Default: A Sentiment Analysis Approach

Hung Ba[a,b], Su Nguyen[c], and Nam Huynh[a]

[a]School of Knowledge Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[b]Business School, The University of Edinburgh, Edinburgh, UK
[c]Business School, La Trobe University, Melbourne, Australia

ARTICLE HISTORY
Compiled August 11, 2019

ABSTRACT

Current works in textual analysis for the financial reports were trying to quantify textual data into meaningful predictors for future financial performance or stock returns. However, little works have been paid to determining predictors for enterprise survival probability using context-rich textual financial reports, especially in discrete-time settings. Our work proposes a new financial sentiment measure on top of the dictionary-based sentiment analysis to determine whether the manager's reflections towards their company's performance, expressed at various aspects, are positive, negative, or neutral in the business and financial context. This research then examines the effects of the proposed measure and other conventional predictors on corporate default. We especially stress on the predictivity of textual analysis using endogenous source for credit risk modelling and how textual-based predictors differ from traditional predictors in terms of timeliness and accurateness. This research is built on a large sample of US enterprises with company fixed characteristics and financial ratios, which is further augmented with financial sentiment. Based on a financial sentiment dictionary, perform dictionary-based sentiment analysis to summarise to what extend textual data could improve corporate default prediction.

KEYWORDS
credit scoring; textual analysis; predictiviy; dictionary-based; sentiment analysis

Email: hungba@jaist.ac.jp—hung.nguyen@ed.ac.uk

1. Introduction

United States Securities and Exchanges Commission (SEC) with the ultimate responsibility of protecting investor, maintaining the market efficiency, and control its capital formation, is having listed US enterprises filing their periodic financial reports via EDGAR (Electronic Data Gathering, Analysis, and Retrieval) system with 10-Ks and 10-Qs forms for annual and quarterly reports, respectively. In these forms, besides their detail financial statements, enterprises need to include their Management Discussion & Analysis (MDA[1]) in the fillings, a forward-looking statement, in which the executives examine their company's performance, address of compliance & risks, and express their views on future goals and new projects. This homogeneous text provides an extremely rich-feature dataset that could be useful for credit risk modelling for listed firms in the US with not just financial statements, but unstructured, textual reports.

On the other hand, most current works in textual analysis on the financial reports were trying to quantify textual data into meaningful predictors for future company financial performance. However, little works have been paid to forming predictors for enterprise survival probability using textual data from the financial reports (for the detailed reviews of financial textual analysis and financial disclosures, reader could refer to Leuz and Wysocki (2016), Loughran and Mcdonald (2016), and Elshandidy et al. (2018)). Yet the predictive power of the new textual-based predictor are also needed to be assessed and compared with traditional predictors in both cross-sectional and longitudinal settings.

This work first create new features from text data of the fillings, it then further confirms that these textual features are significant in predicting corporate default. Based on that, we compare the traditional models built on accounting data with the ones trained on textual data. The results not only confirm the comparable predictive power of textual data but also demonstrate that the combined model built on both types of data could further improve both model fitness and model predictivity power. The experiments are carried under three treatment of imbalanced dataset (IDS), with stratified k-fold cross validation, and using several predictive power measurements. In what follows, we first present the relevant literature on classification models and textual analysis in credit risk in section 2. Section 3 is devoted to the process of mining MDAs, forming textual features, and comparison frameworks. The results are presented in section 4 and we conclude our work in section 5.

2. Literature Review

Single-period or cross-sectional, statistical forecasting models are estimated based on data recorded over a single period of time (Altman, 1968; Altman et al., 2010), and they mostly employ financial ratios as a parametric approach in credit scoring. Recently, machine learning models as better alternatives are being used more frequently as the computer powers have increase significantly. Jo et al. (1997) pointed out that artificial neuron network (ANN) outperformed statistical models. Huang Huang et al. (2004) reaffirmed the performance of support vector machine

[1] https://www.sec.gov/corpfin/cf-manual/topic-9

(SVM) as it is comparable with the back-propagation neuron network and genetic programming. The study of Tsai and Hung (2014) for four approaches in combining the statistical and machine learning models indicated that the combined model of LR and ANN provided the best result. In general, machine learning models are state of the art and can improve the predictive accuracy, however, their practical application is questionable since it lacks the ease of interpretation, explanation, and understanding from the practitioners' viewpoint (Baesens et al., 2003; Sun et al., 2014).

At the same time, there are several works on quantifying the textual data into meaningful predictors for predicting future company financial performance(Healy et al., 1999; Lawrence, 2013), or stock returns (Kraus and Feuerriegel, 2017; Zhou, 2018). In the pioneer work of Loughran and Mcdonald (2011), they built a dictionary with six word lists that offers more accurate tone for financial text compare with the traditional Harvard Dictionary, based on this dictionary, Engelberg et al. (2012) showed that public news provide valuable trading chances for competent information processing short sellers; Bonsall et al. (2017) created a new financial reporting readability measure; Bonsall and Miller (2017) showed that firms with improved readability of their filings have better ratings, lower bond rating disagreement, and lower cost of debt. More recently, Zhou (2018) and Jiang et al. (2019) analysed and examined the relationship of homogeneous and heterogeneous sources of financial texts and indicated that higher manager sentiment followed by lower earnings disruption and higher investment growth. Gandhi et al. (2018) used sentiment words to examine the financial distress of US banks, their empirical findings suggested that more negative words in the reports is related to a higher probability of distressed delisting subsequently. In addition, little works have been devoted to the performance of classification models built on textual features, and relatively compare them with the traditonal ones.

Regarding the employed features for the classifiers, Altman (1968) z-scores are well-known features in building corporate default prediction models. Further extension are also made such as Altman et al. (2010) for Small and Medium Enterprises (SMEs). Different financial institutes might use different feature set for their final models, depend on their risk apatite. Hence, it is desirable to have a comparison for the predictive power of other alternative features built from textual analysis with the traditional features.

In addition, we want to further stress that one crucial problem in the current practices of credit scoring modeling is imbalanced dataset (IDS), built or trained using IDS, classifiers tends to perform extremely poor on minority class despite producing high accuracy measures Chen et al. (2016). The IDS characteristic naturally exists in most of credit portfolios data since the majority of corporate borrowers are good and those bad ones are rare. To mitigate this problem, at data level, researchers could perform undersampling, oversampling, or creating synthetic samples based on SMOTE Chawla et al. (2002) or ROSE Menardi and Torelli (2014) techniques. Despite literature shows that there are sophisticated methods of creating additional data for the minority class, including SMOTE Chawla et al. (2002) or ROSE Menardi and Torelli (2014), however, these methods can not effectively deal with categorical features and might produce unrealistic samples (e.g. companies with -1 employee), hence we proceed to employ undersampling and oversampling at the final stage of modeling process.

By and large, this study uses forward-looking statement, MDAs, in the 10K fillings of US listed firm from 1997 to 2018 to examine how could textual data help predicting the default/liquidation firms.

3. Methodology

3.1. Financial wordlist

Loughran and Mcdonald (2011) showed that some words which are negative in Harvard IV dictionary[2] are actually neutral or even positive in the financial context, such as cancer, depreciation, liability, and so forth. To mitigate the proxy for industry or other unintended effects in using general sentiment dictionary or wordlist, they proposed five word lists: positive; uncertainty; litigious; strong modal words; and weak modal words[3]. These wordlists are related to market reactions such as trading volume, unexpected earnings, and subsequent stock return volatility. We present on what follows how we form textual features using counting and dictionary-based classifier based on these wordlists.

3.2. Dictionary-based sentiment classifier

Dictionary-based (or rule-based) sentiment classifier uses several rules to calculate sentiment score using sentiment words from a lexicon or wordlist, positive words are assigned +1 sentiment scores, while negative and uncertainty words are assigned -1 scores, the sentence sentiment will be calculated based on the aggregate score adjusted for rules. In this study, we employ three simple rules, based on Loughran and Mcdonald (2011) financial wordlists (LM) as follows:

(1) count the number of negative and positive words in a sentence,
(2) consider the shifting of sentiment with constrastive conjunction such as 'but', 'however', 'despite', 'neither' and so forth, and
(3) examine the tri-gram preceding the lexical feature for flipping polarity of the text.

3.3. Textual features

We then form the following textual features:

(1) Percentage of negative, positive, uncertainty, litigious, modal strong, modal weak words in the entire filling, namely pneg, ppos, punc, plit, pmods, and pmodw, respectively. We further compute the sentiment feature (senti) by taking the difference between ppos and pneg.
(2) Count of negative, positive, uncertainty, litigious, modal strong, modal weak words in the MDA section only, namely NEG, POS, UNC, LIT, MODS, and MODW, respectively.
(3) Count of negative, positive, and neutral sentences for each MDA namely CNEG, CPOS, and CNEU, respectively, using dictionary-based classifier defined above. We then further computer the percentage of negative and positive sentences in an MDA as PCNEG and PCPOS.

[2]http://www.wjh.harvard.edu/ inquirer/homecat.htm
[3]Available at http://www.nd.edu/ mcdonald/Word_Lists.html

3.4. Predictive measurement

Having more textual features, we proceed to examine their predictive performance and examine to what extend they could complement the financial features. Regarding the final model performance on predicting default/liquidation firms, it is advisable to have measurements that captures difference aspects of the classifiers, especially with the present of IDS. Goadrich Davis and Goadrich (2006) proposed to used Area Under Precision-Recall Curve (AUPRC) when we perform classification on an IDS and want to concentrate on the positive examples, the default corporates. Specifically, sensitivity is directly influenced by class imbalance, whereas True Positive Rate only depends on positives. In this particular viewpoint of credit risk modeling, we pay more attention to Recall value as in total number of predicted bad corporate loan applications, how many of them are actually bad as bad loans could easily wipe out all the interest profit of entire loan portfolio. Generally, in this study, we employ five metrics to:

- assess the correctness of the models categorical predictions:
  - Precision:

$$P = Pr(Y = 1|\hat{Y} = 1).$$

  - Recall:

$$R = Pr(\hat{Y} = 1|Y = 1).$$

  where 1 denotes a default class, 0 denotes a non default class, and $\hat{Y}$ is the estimate of the true class label Y.
- assess the discriminatory ability of the scorecard:
  - Area Under ROC curve (AUC). This metric is independent with class distribution and shows the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example, i.e.:

$$\text{AUC} = Pr\big(score(x^+) > score(x^-)\big).$$

  - Area Under Precision-Recall Curve (AUPRC) (Davis and Goadrich, 2006). This metric shows how meaningful is a default loan application predicted by the classifier given the baseline probabilities of loan assessment problem, i.e.:

$$\text{AUPRC} = \sum_n (R_n - R_{n-1}) P_n.$$

  where P_n and R_n are the Precision and Recall at the n^{th} threshold respectively.
- assess the accuracy of the scorecards probability predictions: Brier Score (BS), it can be computed using the following formula:

$$BS = \frac{1}{N} \sum_{t=1}^{N} (f_t - o_t)^2.$$

  where f_t is the predicted probability, o_t is the actual outcome of the observation t, and N is the number of observations.

4. Data

4.1. Financial data

We collect accounting data from Wharton Research Data Services (WRDS) for all listed firms in US from 1997-2018, specifically, elements include Total Assets (at), Total Current Assets (act), Intangible Assets (intan), Total Inventory (invt), Cash (ch), Total Receivables (recvt), Total Dividend (dvt), Total Current Liabilities (lct), Working Capital (wcap), Total Revenue (revt), Retain Earning (re), Earning Before Interests and Taxes (ebit), Total Market Value (mkvalt), Common Stock (cstk), Sale (sale), Shareholder Equity (seq), Net Income (ni), and Note-payables (np) are extracted from financial statements. Detail statistics of these financial report elements are presented in Table 5 of the Appendix. Based on these elements, we impute for the missing observations using multiple imputation using chain equations - MICE (Buuren and Groothuis-Oudshoorn, 2011) and calculate the 5-factors z-score (Z1-Z5) as Altman (1968) and 5-factor for SMEs as Altman et al. (2010) (A1-A5) to proxy for the financial accounting features. As for the default flags, firms are marked default if they filed for liquidation under Chapter 7 or Chapter 11 bankruptcy fillings[4].

4.2. 10K filling

Raw filings of listed firms in US could be retrieved from EDGAR (Electronic Data Gathering, Analysis, and Retrieval system[5]). After removing Tables, Figures, attached PDFs, and other redundant elements, we extract the MDA section in each filing. We cover 10K and 10KSB fillings in this study as other types of filling either notice a delay in document fillings (10K405) or a transition of accounting period (10KT and 10K405T).

We present the descriptive statistics of MDA data we mined from SEC fillings in Table 1, other statistics relating to the accounting data, reader could refer to Table 5 in the Appendix. From 1997 to 2018, the number of MDAs we find from the 10Ks fillings ranging from minimum 7092 (2018) to maximum 12,475 (2005), it start at 8711 MDAs found in 1997, increases to 12475 MDAs in 2005, and then starts decreasing down 7092 in 2018. There is a tendency of decreasing number of MDAs since 2008, the originated year of the recent crisis. In general, despite having lower number of fillings for the recent years, the number of sentences in a MDA is increasing in both mean and median.

Default of firms over years is presented in Figure 1 as follows:

[4]https://www.sec.gov/reportspubs/investor-publications/investorpubsbankrupthtm.html
[5]https://www.sec.gov/edgar/aboutedgar.htm

	#MDAs	#Not found	#uncommon	#omitted	#Sentences	mean	median	#tokens
1997	8711	422	7	561	740208	85	64	47328
1998	8931	454	4	563	862791	96.6	73	51372
1999	8934	391	2	611	1044325	116.9	89	54665
2000	9508	458	8	580	1065612	112.1	79	57295
2001	9447	424	1	510	1131682	119.8	83	59454
2002	10179	434	2	806	1488084	146.2	89	64979
2003	11878	419	6	1236	2107816	177.5	117	74079
2004	12124	390	4	1496	2274859	187.6	115	76837
2005	12475	635	2	2016	2501571	200.5	120	81210
2006	12251	678	5	1863	2472559	201.8	135	81163
2007	12087	485	5	1840	2526728	209	145	83147
2008	11432	443	6	1470	2519519	220.4	156	83350
2009	9919	366	4	769	2525077	254.6	189	79620
2010	9165	190	3	676	2405854	262.5	199	78967
2011	8840	162	2	659	2290261	259.1	193	78147
2012	8393	175	1	693	2214756	263.9	195	75322
2013	8105	186	1	677	2183919	269.5	203	74772
2014	8084	184	1	751	2193408	271.3	202	76434
2015	7985	182	1	912	2181273	273.2	204	76066
2016	7589	158	2	1081	2077774	273.8	201	74669
2017	7248	184	1	1113	1931412	266.5	192	71713
2018	7092	256	0	1136	1850778	261	186.5	71617

#MDAs is the total number of MDAs; #Not found is the number of fillings that do not have MDA; #uncommon is the number of uncommon MDAs that we are unable to trace the sections they begin or end with; #omitted is the number of fillings that have the MDA section omitted; #Sentences is the total number of sentences of all MDAs; mean and median are the mean and median of number of sentence in MDAs, respectively; #tokens is the total number of unique words in all MDAs.

Table 1.: MDAs from fillings

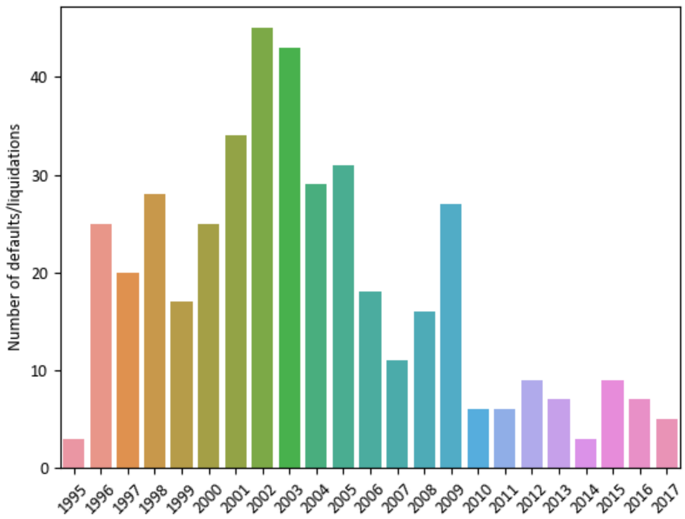

Figure 1.: Number of defaults per year

The number of defaults reaches highest value of 43 firms in 2002 after 2001 financial crisis, we could also observe a spike of default firms in 2009 follows 2008 crisis. We further devise the number of default under each GICS [6], which is an efficient investment

[6]https://www.msci.com/gics

tool to capture the breadth, depth and evolution of industry sectors, in Table 2.

GICS	#Defaults	#Company	%Defaults
10	31	5923	0.52
15	24	5713	0.42
20	36	10723	0.33
25	47	7871	0.59
30	11	3969	0.28
35	21	12662	0.17
40	1	171	0.58
45	19	10260	0.18
50	18	1671	1.07
55	_	56	_
60	_	119	_
Total	208	59163	0.35

Table 2.: GICS sector codes and number of defaults

The sector that has the highest number of defaults is 25-Consumer Discretionary, followed by 20-Industrials with 36 firms. Those smallest sectors, including 60-Real-Estate, 55-Utilities, and 40-Financials having only one default firm (as in the initial stage, we exclude the financial industry in WRDS filter using SIC industry code). This Table show the extremely high degree of imbalance in our data, total number of default/liquidation firms account for just 0.35% of all firms. This suggest, if trained using original data, classifiers might overfitted toward the majority class and the performance measures might not be comparable. At the final stage, we merge the financial data and SEC fillings using CIK, form type, and filling date. The final data consist of 50,957 firms, of which 175 are default (approximate 0.34%).

5. Results

In this section, we present the comparison of regularized logistic regression models fitted for (i) financial features only, (ii) textual features only, and (iii) combine features in Table 3. The performance in term of predictive power are presented in Table 4 where we show the performance of traditional financial model and combined model using five performance metrics under 10-fold stratified cross validation. The results are reported under gridsearch for the best parameter C of the logistic regression classifier and further augmented with undersampling and oversampling to treat the IDS problem.

5.1. Regression

The regression results in Table 3 show three models fitted using logistic regression, control for size and industry, and 1 year lag. Full detail are presented in Table 6 of Appendix. **z-score** is the model fitted using 10 financial features computed as Altman (1968); Altman et al. (2010), **Text** is the model fitted using 18 textual features which are described in subsection 3.3, and the final model, **Combine**, is fitted using both financial and textual features. Regarding the significant of financial features, our results are inline with the current literature of corporate default/liquidation modelling (Altman et al., 2010). On the other hand, the textual features are also significant. As for the entire filling, the percentage of modal-strong and modal-weak words are both

	z-score	Text	Combine
Intercept	-0.1979	-2.3628*	0.9382
	(0.3928)	(1.3917)	(1.4329)
A1	-1.6923***		-1.2690**
	(0.6239)		(0.5974)
A2	-0.1909		-0.1784
	(0.2553)		(0.2514)
A3	-1.3933***		-1.2284***
	(0.3249)		(0.3387)
A4	-2.6194***		-2.5920***
	(0.6297)		(0.6561)
A5	-1.5063***		-1.7554***
	(0.4837)		(0.5053)
Z1	-0.6510		-0.2819
	(0.4196)		(0.4241)
Z2	-0.0290		0.1458
	(0.4953)		(0.5158)
Z3	-1.3672**		-1.2314*
	(0.6338)		(0.6465)
Z4	-1.6061***		-1.5825***
	(0.2964)		(0.3103)
Z5	0.7518**		0.3791
	(0.3037)		(0.3273)
CNEG		3.8558**	4.8806***
		(1.6699)	(1.7392)
CNEU		-1.8360	-2.4985**
		(1.2439)	(1.2530)
CPOS		-0.0269	0.1041
		(1.2551)	(1.2717)
PCNEG		0.3164	-0.5290
		(0.5401)	(0.5831)
PCPOS		-1.0322**	-0.7771
		(0.5259)	(0.5329)
LIT		-0.3925	-0.1917
		(0.9393)	(0.9811)
MODS		0.6659	-0.1615
		(0.7192)	(0.7440)
MODW		-0.5820	-0.8320
		(0.8805)	(0.8848)
NEG		0.4524	-0.7297
		(1.5948)	(1.6397)
POS		-0.4296	0.6396
		(1.2131)	(1.2522)
UNC		-1.1214	-0.3036
		(1.2207)	(1.1968)
plit		0.4959	1.0975**
		(0.4263)	(0.4348)
pmods		1.1790***	0.6101*
		(0.3179)	(0.3391)
pmodw		-1.5003***	-1.9683***
		(0.5581)	(0.5770)
pneg		-1.6440	-1.4065
		(1.5806)	(1.5784)
ppos		0.8212	1.0456
		(0.6882)	(0.6970)
punc		-0.9387	-0.1654
		(0.5879)	(0.6039)
senti		-3.5038**	-2.8255*
		(1.6750)	(1.6700)
N	50957	50957	50957
AIC	2042.7632	2124.7182	1962.9561
BIC	2219.5379	2372.2029	2298.8281
LLR	332.5181	266.5631	448.3253
R-squared	0.1424	0.1141	0.1920

N is the number of observations and LLR is the log-rank test statistics.
***, **, and * indicate 99%, 95% and 90% significant levels, respectively.

Table 3.: Regression

significant, where the higher the percentage of modal-strong words, the more likely the firm to be default. Focusing on MDA section only, we find the more negative sentences in a MDA a firm has, the higher the probability of it becoming default. On the other hand, the higher percentage of positive sentences relative to total number of sentences in MDA, the lower the default probability. As **senti** is the difference between percentage of positive and negative words, this shows that firm having more positive words relative to negative words in their fillings have lower probability of default.

In term of combined model, excluding Z5 and percentage of positive sentences (PCPOS), all the other financial and textual features remain significant. Interestingly, the number of neutral sentences (CNEU) is significant, suggests that, the higher the neutral sentences in the MDA, the lower the likely of the firm to be default. Meanwhile, the higher the number of litigious words in a MDA, the higher the likelihood of firm to be default.

All in all, we demonstrate that, the LM wordlists help not only forming new textual features on explaining default, but also complement effectively with traditional financial-based features. In what follows, we further examine the predictive powers of the combined model with the traditional one.

5.2. Predictive power comparison

In this subsection, the performance of combined model using 28 features is compared with the 10-financial features under no sampling, undersampling and oversampling strategies. We report the 10-fold stratified cross-validation using gridsearch for parameter C of logistic regression classifier.

	z-score			Combine		
	N	U	O	N	U	O
AUC	0.8324	0.8303	0.8326	0.8586***	0.8569***	0.8590***
F1	0.4991	0.4284	0.4327	0.4991	0.4395***	0.4514***
R	0.0000	0.7765	0.7628	0.0000	0.7829	0.7512***
P	0.0000	0.0108	0.0111	0.0000	0.0113***	0.0125***
AUPRC	0.0369	0.0351	0.0286	0.0535***	0.0384	0.0352***
BS	-0.0034	-0.1842	-0.1692	-0.0034	-0.1692***	-0.1448***

N is no sampling, U is undersampling, and O is oversampling strategy.
*** indicates the correspoding repeated measure t-test is significant at 99%.

Table 4.: Predictive power comparison

First, we notice the very poor performance (especially in Recall, Precision, and hence F1 score) of both models on the original data, where we do not employ any balancing strategy on the training set. Also, there is an uplift in AUC and AUPRC using combined model with both financial and textual features. Across all sampling settings, the oversampling produces the best model performance thank to its generating instead of removing observations as in undersampling setting. Compare with traditional model, combined model has higher predictive power in 4 over 6 measures and all 6 measures in undersampling and oversampling strategy, respectively.

6. Conclusions

Textual data are gathering many attentions thank to its complement with other sources of data in explaining the manager sentiment and stock returns (Lopez Lira, 2019), uncovering the role of investment analyst report (Huang et al., 2018), or improving manager sentiment tone understanding (Zhou, 2018). This study further shows that, they improve the traditional model built on the accounting data in predicting corporate default/liquidation.

By using more than 50k observations of listed firms in the US market from 1997 to 2018, and with simple counting for sentiment words both in the entire filings or in the MDA section of the filing, we demonstrate the high predictivity power of textual features in building forecasting models. Despite the severe problem of imbalanced dataset in default/liquidation prediction, the textual features are significant in the regression model and further improve prediction model in almost 5 over 6 predictive measurements.

This study is without its limitation, first, we just examine the US listed firms, which have the benefit of the availability of textual data. Second, assigning -1 and +1 score for positive and negative words are somewhat harsh, since words might have difference level of negative or positive, further work could examine to what degree we should assign negative, positive or neutral to a word or the entire sentence and radiate to the entire MDA or filing. Another avenue could be exploring which aspects of the business are the worst or best performers using sentiment analysis.

Acknowledgement

Hung Ba is supported by JAIST Research Grant and Doctoral Research Fellow Grant No. 238003.

References

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, 23(4):589–609.

Altman, E. I., Sabato, G., and Wilson, N. (2010). The value of non-financial information in SME risk management. *Journal of Credit Risk*, 6(2):95–127.

Baesens, B., Gestel, T. V., Viaene, S., Stepanova, M., Suykens, J., and Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6):627–635.

Bonsall, S. B., Leone, A. J., Miller, B. P., and Rennekamp, K. (2017). A plain English measure of financial reporting readability. *Journal of Accounting and Economics*, 63(2-3):329–357.

Bonsall, S. B. and Miller, B. P. (2017). The impact of narrative disclosure readability on bond ratings and the cost of debt. *Rev Account Stud*, 22(2):608–643.

Buuren, S. v. and Groothuis-Oudshoorn, K. (2011). **mice**: Multivariate Imputation by Chained Equations in *R*. *Journal of Statistical Software*, 45(3).

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chen, N., Ribeiro, B., and Chen, A. (2016). Financial credit risk assessment: a recent review. *Artificial Intelligence Review*, 45(1):1–23.

Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM.

Elshandidy, T., Shrives, P. J., Bamber, M., and Abraham, S. (2018). Risk reporting: A review of the literature and implications for future research. *Journal of Accounting Literature*, 40:54–82.

Engelberg, J. E., Reed, A. V., and Ringgenberg, M. C. (2012). How are shorts informed? *Journal of Financial Economics*, 105(2):260–278.

Gandhi, P., Loughran, T., and McDonald, B. (2018). Using Annual Report Sentiment as a Proxy for Financial Distress in U.S. Banks. SSRN Scholarly Paper ID 2905225, Social Science Research Network, Rochester, NY.

Healy, P. M., Hutton, A. P., and Palepu, K. G. (1999). Stock Performance and Intermediation Changes Surrounding Sustained Increases in Disclosure. *Contemporary Accounting Research*, 16(3):485–520.

Huang, A. H., Lehavy, R., Zang, A. Y., and Zheng, R. (2018). Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach. *Management Science*, 64(6):2833–2855.

Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., and Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, 37(4):543–558.

Jiang, F., Lee, J., Martin, X., and Zhou, G. (2019). Manager sentiment and stock returns. *Journal of Financial Economics*, 132(1):126–149.

Jo, H., Han, I., and Lee, H. (1997). Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis. *Expert Systems with Applications*, 13(2):97–108.

Kraus, M. and Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*, 104:38–48.

Lawrence, A. (2013). Individual investors and financial disclosure. *Journal of Accounting and Economics*, 56(1):130–147.

Leuz, C. and Wysocki, P. D. (2016). The Economics of Disclosure and Financial Reporting Regulation: Evidence and Suggestions for Future Research: Disclosure and Financial Reporting Regulation. *Journal of Accounting Research*, 54(2):525–622.

Lopez Lira, A. (2019). Risk Factors That Matter: Textual Analysis of Risk Disclosures for the Cross-Section of Returns. *SSRN Journal*.

Loughran, T. and Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Loughran, T. and Mcdonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, 54(4):1187–1230.

Menardi, G. and Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1):92–122.

Sun, J., Li, H., Huang, Q.-H., and He, K.-Y. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems*, 57:41–56.

Tsai, C.-F. and Hung, C. (2014). Modeling credit scoring using neural network ensembles. *Kybernetes*, 43(7):1114–1123.

Zhou, G. (2018). Measuring Investor Sentiment. *Annual Review of Financial Economics*, 10:239–259.

Appendix

PANEL 1: Financial statements elements statistics

	at	act	intan	invt	ch	dvt	lct	lt	wcap	revt
count	86196.000	84797.000	81638.000	85708.000	85565.000	84619.000	84837.000	86196.000	84731.000	85921.000
mean	4118.302	1217.743	858.391	284.317	272.502	89.814	928.688	2491.773	279.546	2934.885
std	19224.416	5655.499	5462.940	1347.345	1292.234	592.195	4827.252	12177.598	1799.280	14150.218
min	0.001	-0.168	-0.423	0.000	-0.134	-325.377	0.000	0.001	-43132.545	-1964.999
25%	26.927	14.269	0.000	0.381	1.805	0.000	5.922	9.265	1.703	13.857
50%	186.378	82.971	3.870	10.634	13.877	0.000	33.375	72.032	28.528	138.835
75%	1246.340	414.989	116.742	92.331	82.579	3.874	213.332	687.642	155.221	969.236
max	507560.425	192486.646	225278.000	48586.913	53528.000	67643.805	192819.656	460442.000	56120.000	470171.000

	re	ebit	mkvalt	sale	seq	ni	dltt	dm	emp	emp	gdwl
count	84042.000	85793.000	61887.000	85807.000	86195.000	85805.000	86141.000	80436.000	78807.000	78807.000	80728.000
mean	821.767	333.038	3064.844	2937.978	1554.264	176.908	921.058	150.610	8.641	8.641	498.611
std	7845.878	1800.682	15901.855	14159.247	8001.189	1527.857	4476.683	1075.271	30.442	30.442	3138.554
min	-143336.328	-25913.000	0.000	-1964.999	-86154.000	-98696.000	-0.023	0.000	0.000	0.000	-0.423
25%	-58.985	-3.571	27.282	13.832	8.472	-8.340	0.000	0.000	0.088	0.088	0.000
50%	-1.867	4.762	173.450	138.825	74.588	0.715	8.898	0.180	0.619	0.619	0.000
75%	122.538	84.664	989.844	969.980	463.727	37.491	287.779	19.804	4.000	4.000	55.402
max	398278.000	71230.000	790050.098	470171.000	284434.000	104821.000	207174.000	59127.799	863.824	863.824	146583.307

PANEL 2: Demographic elements statistics

	addzip	city	state	county	sic	ggroup	gind	gsector	gsubind	idbflag	incorp	spcsrc	au	auop
count	86196	86196	86196	86196	86196	86196	86196	86196	86196	86196	86196	86196	86196	86196
unique	4601	2178	62	37	266	25	67	12	148	2	55	10	25	7
top	nan	Houston	nan	nan	2836	3520.0	352010.0	35.0	35201010.0	D	DE	nan	9.0	1.0
freq	1376	3302	13037	86026	5422	11224	6770	17232	6770	72957	45452	37054	17463	57263

Table 5.: Descriptive statistics for accounting data

	z-score	Text	Combine
Intercept	-0.1979	-2.3628*	0.9382
	(0.3928)	(1.3917)	(1.4329)
A1	-1.6923***		-1.2690**
	(0.6239)		(0.5974)
A2	-0.1909		-0.1784
	(0.2553)		(0.2514)
A3	-1.3933***		-1.2284***
	(0.3249)		(0.3387)
A4	-2.6194***		-2.5920***
	(0.6297)		(0.6561)
A5	-1.5063***		-1.7554***
	(0.4837)		(0.5053)
Z1	-0.6510		-0.2819
	(0.4196)		(0.4241)
Z2	-0.0290		0.1458
	(0.4953)		(0.5158)
Z3	-1.3672**		-1.2314*
	(0.6338)		(0.6465)
Z4	-1.6061***		-1.5825***
	(0.2964)		(0.3103)
Z5	0.7518**		0.3791
	(0.3037)		(0.3273)
CNEG		3.8558**	4.8806***
		(1.6699)	(1.7392)
CNEU		-1.8360	-2.4985**
		(1.2439)	(1.2530)
CPOS		-0.0269	0.1041
		(1.2551)	(1.2717)
LIT		-0.3925	-0.1917
		(0.9393)	(0.9811)
M1		0.6659	-0.1615
		(0.7192)	(0.7440)
M3		-0.5820	-0.8320
		(0.8805)	(0.8848)
NEG		0.4524	-0.7297
		(1.5948)	(1.6397)
POS		-0.4296	0.6396
		(1.2131)	(1.2522)
UNC		-1.1214	-0.3036
		(1.2207)	(1.1968)
pcneg		0.3164	-0.5290
		(0.5401)	(0.5831)
pcpos		-1.0322**	-0.7771
		(0.5259)	(0.5329)
plit		0.4959	1.0975**
		(0.4263)	(0.4348)
pmods		1.1790***	0.6101*
		(0.3179)	(0.3391)
pmodw		-1.5003***	-1.9683***
		(0.5581)	(0.5770)
pneg		-1.6440	-1.4065
		(1.5806)	(1.5784)
ppos		0.8212	1.0456
		(0.6882)	(0.6970)
punc		-0.9387	-0.1654
		(0.5879)	(0.6039)
senti		-3.5038**	-2.8255*
		(1.6750)	(1.6700)
Size	-1.0784***	-0.4262**	-0.8799***
	(0.1981)	(0.1871)	(0.2091)
G10	-0.5093	-0.4919	-0.2778
	(0.3184)	(0.3203)	(0.3334)
G15	-0.7189**	-0.9124***	-0.7458**
	(0.3258)	(0.3336)	(0.3392)
G20	-0.8731***	-1.2089***	-0.8295**
	(0.3159)	(0.3206)	(0.3258)
G25	-0.4067	-0.6603**	-0.4146
	(0.2996)	(0.3063)	(0.3104)
G30	-0.8202**	-0.8605**	-0.6102
	(0.3958)	(0.4002)	(0.4071)
G35	-1.9317***	-1.9622***	-1.7328***
	(0.3911)	(0.3857)	(0.4007)
G40	-0.0136	-0.3324	0.0884
	(1.0552)	(1.0472)	(1.0640)
G45	-1.1523***	-1.8093***	-1.1684***
	(0.3736)	(0.3692)	(0.3839)
N	50957.0000	50957.0000	50957.0000
AIC	2042.7632	2124.7182	1962.9561
BIC	2219.5379	2372.2029	2298.8281
LLR	332.5181	266.5631	448.3253
R-squared	0.1424	0.1141	0.1920

Table 6.: Regression - All variables